THE UNIVERSITY
of EDINBURGH

Identification of functional single nucleotide polymorphisms (SNPs) in High Risk-Human Papillomavirus (HR-HPV) related diseases

Duanduan Cong

从端端

Doctor of Philosophy

The University of Edinburgh

2017

THE UNIVERSITY *of* EDINBURGH

Declaration

I hereby declare that this thesis has been composed by myself, describes my own work, and has not been submitted for any other degree or professional qualification. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

September 2017

Acknowledgments

At the beginning of my thesis, I would like to express my sincere gratitude to all those people who have helped me in the past four years. Without their encouragement, constant suggestions and continued support, this thesis would hardly have been completed.

First of all, I would like to show my heartfelt gratitude to my primary supervisor Prof Sarah Howie and my co-supervisors Prof Jurgen Haas and Dr Kate Cuschieri for their invaluable advice, enthusiastic encouragement, patient guidance and supervision for my work. It was a great pleasure working with them. I have enjoyed the opportunities to watch and learn from their knowledge and experience.

In particular:

Sarah, I can't thank you enough for everything that you have taught me. You are the best supervisor I have ever had. You trained me up from scratch. I had very limited lab experience when I joined your group. But now I can step into the research field and be confident about who I am and what I can achieve. Words can neither qualify nor quantify how helpful your guidance and advice have been. You let me know where I want to go in my career and you are one of the reasons why I am always trying my best. You are truly an awesome scientist and a role model for me. I really hope that I can help someone else in the future just like you have helped me.

Jurgen, thank you so much for introducing me into the field of genetic research and offering me the opportunities to learn different kinds of techniques. And also thank you for creating an inspirational environment for students. I always admire your

enthusiasms in scientific work and your countless new ideas and ways of thinking. You set a great model for me. It's so inspiring to be a member of your group.

And Kate, thank you for being an excellent supervisor and a great inspiration. I have learnt a lot from you and I was deeply inspired by your hardworking and passionate attitude. Also thank you so much for being there as my perfect guide to shape my career. Because of you and Sarah, I now firmly believe that female professionals can contribute in science equally or even better than males.

Next, I would like to thank the entire HPV group for the help and support during my whole PhD project. You are all wonderful friends and I am incredibly fortunate to have had the opportunity to work with you. Sincere thanks to Dr Ramya Bhatia for providing me a lot of useful suggestions and helping me through some tough times. Also thanks to my colleagues Dr Michelle Etherson and Luhan Jiang for the assistance with the validation experiment and the chemotaxis assay. Itzi, Elia and Chara, thanks for being there for me when I really needed you. I had a super time with you girls. You don't even know how much your help meant to me. June and Sharon, thank you both for your technical support and also thanks for making the office a fun place to be.

Besides, I am greatly indebted to Dr Rui Chen and Dr Samantha Griffiths for taking the time to provide me initial training in methodology and informative suggestions. I would like to thank Dr Lee Murphy and Mr William Hawkins from Genetic core WTCRF for their help with the SNP array. My grateful thanks are also extended to Prof Colin Semple, Dr Maria Timofeeva and Miss Cat Graham for their advice in statistical analysis, to Mrs Shonna Johnston, Mr Will Ramsay and Dr Mari Pattison

for providing me flow cytometry training and cell sorting service, to Dr Donald Davidson and Dr Brian McHugh for the valuable advice on my project and for letting me use the digital microscope in their lab.

I am also indebted to Scottish HPV archive and Lothian NRS Bioresource for processing and providing precious clinical samples for this project and to the China Scholarship Council for their financial support provided through the China Scholarship Council/University of Edinburgh Joint Scholarship.

In addition, I would like to thank Dr Xiaozhong Zheng, Mrs Ailiang Zhang, Dr Cunjing Yu, Dr Xiaochao Chen, Dr Bingjie Wang and Dr Feng Li for their kindness and support. Thank you all for everything that you have done to help me. I will be forever grateful. I would also like to thank my favourite idol Jurina Matsui for being a constant source of inspiration and encouragement.

And finally I would like to offer my special thanks to my family: my parents and my grandfather for their continuous and unparalleled love and support. You are always the biggest motivator in my life.

Abstract

Persistent infection of the cervix with high risk (HR) types of Human Papilloma Virus (HPV) (HR-HPV) can result in precancerous lesions and cancers. However, most HPV infections can be cleared naturally by the immune response without causing disease. Although genetic variations have long been considered as the main explanation for individual heterogeneity in cancer susceptibility, the underlying mechanisms remain unclear.

In this project, a panel of routinely taken clinical samples was assessed for 32 rationally selected SNPs with allele frequency related to disease outcome using the Taqman® OpenArray® system. The panel incorporated 475 HR-HPV negative, cytologically-normal cervical samples, 413 HR-HPV positive cervical high grade squamous intraepithelial lesion (HSIL) cases and 62 HR-HPV positive cervical cancers. Two SNPs, rs2234671 and rs2623047, were found with significant differences between HR-HPV negative, cytologically-normal samples and HR-HPV positive cervical HSIL cases. In the validation step, these two SNPs were further genotyped in the same set of samples using TaqMan® SNP genotyping assay and/or LightSNiP assay and in additional samples including 83 HR-HPV positive, cytologically-normal cervical samples, 21 HR-HPV positive cervical cancer cases, 129 HR-HPV positive vulval intraepithelial neoplasia cases and 23 HR-HPV positive vulval cancer cases. Statistical analysis was then performed based on pooled and re-grouped genotyping data of the above-mentioned samples under different genetic models so as to evaluate the associations with different stages in the disease process. After validation, SULF1 rs2623047 revealed a strong significant association with the

susceptibility to HR-HPV infection but not with the development of high-grade squamous intraepithelial lesion and the progression to cervical cancer. CXCR1 rs2234671, by contrast, was associated with the progression of HR-HPV-related cancers and the minor allele CXCR1 827C was significantly enriched in HPV16 positive cancers.

CXCR1 is a receptor for the chemokine CXCL8/IL-8 and CXCR1 rs2234671 leads to a serine to threonine change in an extracellular loop of the receptor. Functionally, the CXCR1 827C allele was shown to enhance cell motility in response to IL-8 stimulation in a chemotaxis assay with transiently transfected fibroblasts (HEK293 cells) and also in a wound healing assay with stably transduced cervical cancer (CaSki) cells. In addition, significantly increased cell proliferation upon IL-8 treatment was observed in two cervical cancer derived cell lines, CaSki and SiHa, transduced with CXCR1-827C allele, but not in their CXCR1 827G transduced counterparts.

These findings suggest that SULF1 rs2623047 and CXCR1 rs2234671 may be genetic risk factors for HR-HPV-related cervical disease and CXCR1 rs2234671 might affect HR-HPV-related cancer susceptibility by functionally altering IL-8-CXCR1 signalling. This information has potential for use in the risk stratification of HR-HPV infected women and may also suggest new therapeutic targets to be exploited for treatment of cervical cancer patients.

Lay summary

Human papillomavirus (HPV) is the most common viral infection in the reproductive tract. Most people will get HPV at least once in their lifetime. To date, over 200 different HPV types have been identified. Among these 200 HPV types, at least 14 are classified as high risk HPV (HR-HPV) types because persistent infection with them is necessary for the development of cervical cancer. Although HR-HPV infections are quite prevalent in the population, most of the time they can be cleared by our body's defence systems without any kind of treatment. Only a small number of women will have persistent infection and even fewer will develop cervical cancer. Evidence shows that genetic factors contribute greatly to the increased cancer risk in HR-HPV infected women, but very few of these genetic factors have been identified and the underlying mechanisms remain unclear.

Therefore, this project aimed to find genetic factors that are associated with HR-HPV-related cervical disease and to explain the mechanisms by which they may contribute to the disease risk. Single nucleotide polymorphism (SNP) was selected as the main focus for this project because it is the most common type of genetic difference in whole populations and has frequently been linked to diseases. In the discovery phase, 32 carefully selected SNPs have been tested in a panel of routinely taken clinical cervical samples consisting of 475 HR-HPV negative, abnormal cells-free controls and 475 cases with HR-HPV infections and high-grade abnormalities in their cervix. By comparing the differences between the controls and cases in the genetic make-up of each SNP tested, two novel SNPs, CXCR1 rs2234671 and SULF1 rs2623047 were found to be associated with the risk of

HR-HPV-related cervical disease. After validating the results in the discovery phase using other genotyping methods and testing the two SNPs in additional samples reflecting different stages of the HR-HPV-related cervical disease, SULF1 rs2623047 was found to be associated with the acquisition of the HPV infection whilst CXCR1 rs2234671 was linked to the development of cervical cancer in the presence of a common HR-HPV type, HPV16.

The CXCR1 rs2234671 leads to a structural change in the cell surface molecule CXCR1, which is a receptor for a protein called IL-8. CXCR1 can regulate a wide range of cell functions in the presence of IL-8, such as cell movement and cell growth. In subsequent functional experiments, the serine to threonine change in CXCR1 rs2234671 was found to increase cell movement and growth of HPV16 positive cervical cancer cells in response to IL-8. These findings provided possible functional explanations for the observed association with high grade cervical disease.

Overall, my data reveal two novel SNPs that are associated with different transition states of the natural history of HR-HPV-related cervical disease. Of the two SNPs, CXCR1 rs2234671 was found to affect the development of cervical cancer by changing cell functions mediated by CXCR1 in response to IL-8. These findings provide novel insights into the roles of SNPs in the pathogenesis of HR-HPV-related disease. The associations still need to be further verified in a larger sample set. Once verified, these SNPs may be used as biomarkers, aiding in the advancement of diagnostics and therapeutics.

Contents

List of Figures

List of tables

List of Abbreviations

AIC	Akaike's Information Criterion
APC	Allophycocyanin
APN	Acute pyelonephritis
ATM	Ataxia telangiectasia mutated
ATR	Ataxia telangiectasia and Rad3-related
BIC	Bayesian Information Criterion
BRIP1	BRCA1-interacting protein 1
BSA	Bovine Serum Albumin
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CFSE	5(6)-Carboxyfluorescein diacetate N-succinimidyl ester
CI	Confidence intervals
CIN	Cervical intraepithelial neoplasia
CT	Cycle threshold
CXCR1	C-X-C Motif Chemokine Receptor 1
CYP1A1	Cytochrome P450, family 1, subfamily A, polypeptide 1
CYP2D6	Cytochrome P450 2D6
ddH2O	double distilled H2O
DDR	DNA damage response
DMC1	DNA Meiotic Recombinase 1
DMEM	Dulbecco's Modified Eagle's medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid

DUT	Deoxyuridine Triphosphatase
EDTA	Ethylenediaminetetraacetic acid
eQTL	Expression quantitative trait loci
FBS	Fetal bovine serum
FDR	False discovery rate
FFPE	Formalin-fixed, paraffin-embedded
FIGO	International Federation of Gynaecology and Obstetrics
FSC	Forward light scatter
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
gDNA	genomic DNA
GFP	Green fluorescent protein
GMAF	Global minor allele frequency
GPCR	G protein–coupled receptor
GTF2H4	General Transcription Factor IIH Subunit 4
GWA	Genome wide association
GWAS	Genome wide association study
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HPV	Human papillomavirus
HR-HPV	High risk-Human papillomavirus
HSIL	High grade squamous intraepithelial lesion
HSPGs	Heparan sulfate proteoglycans
HWE	Hardy-Weinberg Equilibrium
IFNG	Interferon Gamma

ISD	Information Services Division
LB	Lysogeny broth
LBC	Liquid-based cytology
LD	Linkage disequilibrium
LOD	Log-odds
LRT	Likelihood ratio test
LSIL	Low grade squamous intraepithelial lesion
MAF	Minor allele frequency
MDSCs	Myeloid-derived suppressor cells
MFI	Mean fluorescence intensity
MGC	Mammalian Gene Collection
MHC	Major histocompatibility complex
MMP-9	Matrix metallopeptidase 9
MTHFR	Methylene tetrahydrofolate reductase
NCBI	National Centre for Biotechnology Information
NHS	National Health Service
OAS3	2'-5'-Oligoadenylate Synthetase 3
OD	Optical Density
ORF	Open reading frame
ORF	Odds ratios
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PRDX3	Peroxiredoxin 3
RT-qPCR	Quantitative reverse transcription PCR

RNA	Ribonucleic acid
RPS19	Ribosomal Protein S19
RPMI	Roswell Park Memorial Institute
RT	Room temperature
SNP	Single nucleotide polymorphism
SSC	Side light scatter
STAT3	Signal transducer and activator of transcription 3
SULF1	Sulfatase 1
TCGA	The Cancer Genome Atlas
TLR9	Toll-like receptor 9
Tm	Melting temperature
TMC6	Transmembrane Channel Like 6
TMC8	Transmembrane Channel Like 8
TP53	Tumour protein p53
VEGF	vascular endothelial growth factor
VEGFA	Vascular endothelial growth factor A
VIN	Vulval intraepithelial neoplasia
WES	whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization
WTCRF	Wellcome Trust Clinical Research Facility
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1

Chapter 1
Introduction

1.1 Human papillomavirus (HPV) infection

1.1.1 Basic structure of human papillomavirus

HPVs are a group of small, non-enveloped circular double-stranded deoxyribonucleic acid (DNA) viruses belonging to the papillomavirus family, which infect cutaneous or mucosal epithelium. The human papillomavirus family is split into 5 recognised genera (alpha, beta, gamma, mu and nu) and a group of unclassified types depending upon their genetic sequence homology (see Figure 1). The prototypical structure of HPV consists of a genome of about 8 kilobase pairs (kb) and a symmetrical icosahedral capsid. The HPV genome contains three functional regions: the long control region, the early coding region and the late coding region. The eight known primary viral protiens of HPV, including the six early proteins (E1, E2, E4, E5, E6 and E7) and the two late proteins (L1 and L2), are encoded by the early region and the late region respectively (Xue et al., 2012c).

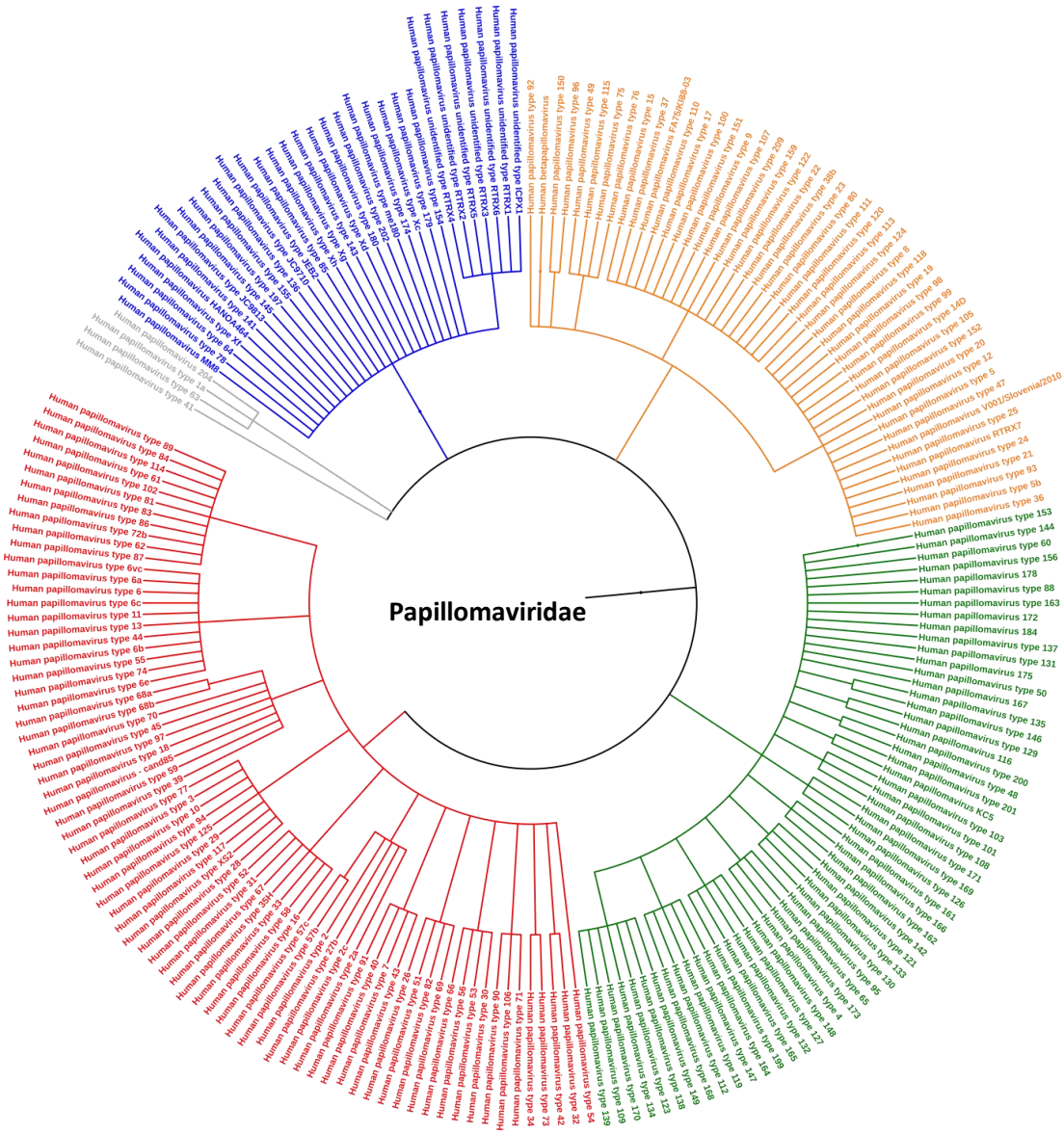


Figure 1 Phylogenetic tree of HPVs

The figure shows the phylogeny of fully sequenced HPVs in the NCBI taxonomy database (released at 07/08/2017) with all alpha genus HPVs in Red, beta genus HPV in orange , gamma genus HPVs in green, mu and nu genus HPVs in grey and other unclassified HPVs in blue. The figure was generated on the basis of Papillomaviridae subtree (NCBI Taxonomy ID: 151340) using online tool phyloT and Interactive Tree Of Life(iTOL) v3(Letunic and Bork, 2016).

L1 and L2 are the structural proteins forming the capsid of the virus. The major capsid protein L1 was found to be responsible for the primary attachment of the virus

to the host cells while the minor capsid protein L2 was reported to be necessary for the genome encapsidation and endosomal escape (Horvath et al., 2010; Xue et al., 2012c). E1 and E2 play essential roles in the initiation of viral replication. By serving as a loader through protein-protein interactions, E2 recruits E1 to the viral repliaction origin. The helicase domain of E1 then unwinds double-stranded DNA and faciliates the replication. Besides, E2 can also act as a tethering factor for the maintaince of the viral genome and controls viral transcription in a dose dependent manner (McBride, 2013). The most commonly investigated E4 is actually a fusion protein E1^E4, which was shown to be important for the amplification of the viral genome as well as the assembly and release of virus (Doorbar, 2013). Unlike the previously mentioned viral proteins, E5 is not a universal protein for all HPV types and is present mainly in alpha genus HPV types. Several subgroups of E5 exist and the alpha group E5, which contains a relatively conserved sequence, was shown to be oncogenic (Müller et al., 2015). E6 and E7 proteins are the two most widely investigated viral proteins among all the primary viral proteins. They are oncoproteins in certain HPV types and faciliate tumourigenesis though a wide variety of mechanisms, of which the most well-characterized include E6 induced degradation of the tumour suppressor protein p53 and E7 induced degradation of retinoblastoma protein (Xue et al., 2012c). Apart from these primary viral proteins, functional isoforms of the primary viral proteins and fusion proteins were also shown to be present in specific HPV types (Graham, 2010). The long control region does not encode any viral proteins, instead, it exerts regulatory functions on viral replication and transcription (Xue et al., 2012a).

1.1.2 Classification of HPVs

The major capsid protein L1 open reading frame (ORF) sequence was found to be the most conserved region within the papillomavirus genome. Therefore, L1 has been chosen as the basis for taxonomy. According to the consensus criteria, a novel HPV type could be identified if its genetic sequence of L1 shares no more than 90% nucleotide sequence identity with all other known HPV types (De Villiers et al., 2004). To date, over 200 types of HPV have been identified and numbered based on this criteria (Van Doorslaer et al., 2013). HPV types that share more than 70% but less than 90% nucleotide sequence identity within the L1 ORF could be grouped together as a species. And an order higher, HPV types that share between 60% to 70% homology of the L1 ORF sequence could be classified as a genus. All the recorded HPV types in the current National Centre for Biotechnology Information (NCBI) taxonomy database are shown in Figure 1.

At least 14 mucosal HPV types in the alpha genus are consensually classified as high risk (oncogenic) types due to their causal associations with cervical cancer, including HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 and 68. Certain HPV types from the beta and gamma genera, such as HPV5, HPV8 and HPV60, have recently attracted more attention since they were found to act as “accomplices” in non-melanoma skin cancer and head and neck cancers (Agalliu et al., 2016; Paradisi et al., 2011). In contrast, mu and nu genera have only been observed in cutaneous infections and have not been associated with malignant diseases (Tommasino, 2014).

1.1.3 Prevalence of HPV infection

HPVs are the most common viral infection in the genital tract. According to a newly released report from National Centre for Health Statistics in the USA, the prevalence of genital HPV in American adults aged 15-59 years were 45.2% and 39.9 % in men and women respectively during 2013-2014. The corresponding data for HR-HPV types were 25.1% and 20.4% (McQuillan et al., 2017). In the UK, HPV was observed in 10.3% of women with normal cervical cytology. In women with abnormal cervical cytology, HPV16 and HPV18, the two most common high risk-HPV (HR-HPV) types, were shown to account for nearly 60% of high-grade cervical intraepithelial lesions and 79% of cervical cancers (Bruni et al., 2017).

Taking advantage of the self-assembly feature of the major capsid protein L1, virus-like particle based prophylactic vaccines have been developed and commercialized to prevent cervical HPV infection. Currently, three HPV vaccines are available on the market, which are Cervarix, Gardasil® and Gardasil®9 respectively. Cervarix is designed to protect against the two most common HR-HPV types HPV16 and HPV18. Although initially targeting only two HR-HPV types, a cross-protection was observed for Cervarix against HPV31, HPV33 and HPV45 (Cuschieri et al., 2016; Paavonen et al., 2009). Apart from HPV16 and HPV18, Gardasil® also protects against HPV6 and HPV11, which cause genital warts. And on the basis of Gardasil®, Gardasil®9 offers extra protection against HPV31, HPV33, HPV45, HPV52 and HPV58. With the widespread introduction of the vaccines, prevalence of HR-HPVs in countries and areas with high vaccine uptake started to exhibit significant changes. As an example, Scotland commenced national HPV

immunisation programme in 2008 and has maintained relatively high vaccination coverage in the applicable populations during these years. A recent study from Scotland showed a significant decline of the prevalence of HPV16 and HPV18 in fully vaccinated women in comparison with their unvaccinated counterparts. Besides, cross protection against several other HR-HPV types and herd protection have also been reported (Cameron et al., 2016). As vaccination continues to gain ground worldwide, it is expected to see more changes in HPV type prevalence in the near future.

1.1.4 Life cycle of HPV and pathogenesis

HPVs display tropism towards multi-layered stratified squamous epithelium (see Figure 2 for a schematic diagram of the HPV life cycle). Normally, the life cycle starts with the infection of dividing basal cells. Through microwounds present in the stratified epithelium, HPV reaches to the basal lamina and enters the target cells with the assistance from various extracellular matrix and cell surface molecules. Among these molecules, heparan sulfate proteoglycans (HSPGs) were considered to be the primary attachment factors for many HPV types (Aksoy et al., 2017). After internalization into basal cells, the capsid of the virus is disassembled in acidic endocytic vesicles. Chaperoned by L2, the viral genome then egresses from endosomes, traffics through the trans-Golgi network and finally reaches nucleus (DiGiuseppe et al., 2017). At the early stage of infection, the viral genome is amplified slowly along with the cell cycle and is subsequently maintained in the basal cells as episomes at low copy numbers. E1 and E2 play essential roles in this stage, though both are expressed at a relatively low level. Expression of E6 and E7

are also observed at this stage but divergence presents between high-risk and low-risk HPV types. E6 and E7 of the high-risk HPV types can promote cell proliferation within the basal and parabasal layers and results in the development of neoplasia (Doorbar et al., 2015). A productive life cycle of the virus continues as infected daughter cells of the asymmetrically dividing basal cells leave the basal layers and start differentiation. E6 and E7 from both high-risk and low-risk HPV types interact with multiple cellular molecules and consequently induce cell cycle re-entering of the differentiated cells in the mid and the upper cell layers, which lays the foundation for the sharp increase in viral copy number during this phase. Expression of other regulatory viral early genes also increases in these cells following activation of corresponding promoters and this directly or indirectly contributes further to the amplification of viral genomes. To complete the full life cycle, the synthetic strategy of the virus switches from early proteins to late proteins while the infected cells finally exit cell cycle and migrate to the upper layers of the stratified epithelium. By altering the splicing pattern, both L1 and L2 are increasingly expressed in the cell cytoplasm and then delivered to the nucleus for virus packaging (Doorbar et al., 2015). After a complicated process of encapsidation and maturation, infectious virions are released from the infected cells with the aid of viral protein E4. This release process disturbs the normal structures of the superficial epithelial layers (Xue et al., 2012b).

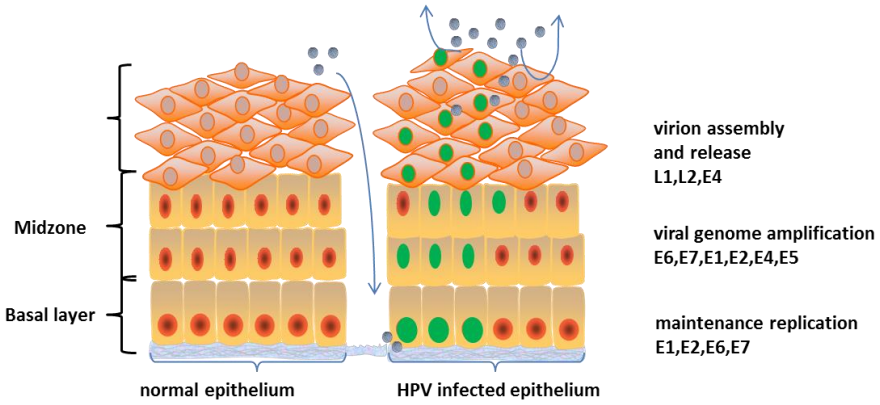


Figure 2 Schematic diagram representing the life cycle of HPV

HPV viruses reach to the basal layer and enter the target cells through a micro-wound or micro-abrasion. The viral life cycle is tightly dependent on the differentiation process of the host epithelial cells. Ordered viral gene expression and life cycle events are indicated on the right side. HPV infected cells are showing as cells with green nuclei while HPV viruses are showing as grey particles.

Intraepithelial lesions caused by HR-HPV types are characterised by increased mitotic activity, nuclear atypicality, cellular immaturity and disorganization and preferentially occur at the transformation zone of the cervix. In low-grade intraepithelial lesions, HR-HPV types go through the above-mentioned productive life cycle and ordered viral gene expression remains unaffected. However, in high-grade intraepithelial lesions, the productive life cycle is retarded due to the deregulation of viral oncoproteins, especially E6 and E7. Deregulated E6 and E7 not only promote cell proliferation throughout the epithelium but also predispose cells to undergo genomic changes or even malignant transformation (Doorbar et al., 2015).

Previous studies found that increased levels of E6 and E7 could cause both numerical and structural chromosomal aberrations by interacting with various cellular components (Graham, 2017; Moody and Laimins, 2010). For example, high risk

HPV E7 could cause aberrant centrosome duplication and subsequently lead to the development of aneuploidy and chromosome missegregation. The centrosome abnormalities have been shown to be mediated through the binding of E7 to γ-tubulin or the retinoblastoma family of proteins. In coordination with E7, E6 could induce centrosome abnormalities by compromising mitotic checkpoint functions through a p53 dependent mechanism (Moody and Laimins, 2010). Also, E6 and E7 could cause DNA damage indirectly through the induction of oxidative stress (Marullo et al., 2015).

To counteract naturally occurring or exogenous agents-induced DNA damage, cells have evolved a complex network of pathways collectively named as DNA damage response (DDR). Upon sensing DNA damage, DDR pathways are activated, leading to cell cycle arrest, DNA repair and cell death (Hollingworth and Grand, 2015). In HPV infection, the major regulators of the DDR pathways, such as ATM (Ataxia telangiectasia mutated) and ATR (Ataxia telangiectasia and Rad3-related) kinases, are activated and hijacked by HPV to support its genome maintenance and facilitate productive viral replication (Anacker and Moody, 2017). Meanwhile, E6 and E7 play an important role in countering downstream of the DDR cascades and consequently allow accumulation of DNA damage within infected cells. For example, both E6 and E7 have been shown to be capable of abrogating p53-dependent cell cycle checkpoints (Fischer et al., 2017). E7 has also been reported to interrupt cell cycle checkpoints through interactions with histone deacetylases, cyclins and cyclin-dependent kinase inhibitors (Moody and Laimins, 2010). Additionally, E6 could disrupt postmitotic checkpoint through a p53 independent mechanism by upregulating cyclin-dependent kinase 1 (Zhang et al., 2015).

In conjunction with the abrogation of cell cycle checkpoints, E6 and E7 could disrupt normal DNA repair processes and prevent infected cells from undergoing apoptosis and senescence. Attenuation of single strand break repair was found to be mediated by E6 through its interaction with O (6)-methylguanine-DNA methyltransferase and protein XRCC1 (Iftner et al., 2002; Srivenugopal and Ali-Osman, 2002). Base excision repair could also be attenuated through the interaction between E6 and XRCC1. In addition, both E6 and E7 have been shown to impede double strand break repair by impairing homologous recombination pathway (Wallace et al., 2017). Furthermore, E6 and E7 could jointly promote the activities of several DNA methyltransferases, thereby resulting in epigenetic inactivation of tumour suppressor and immune genes (Soto et al., 2017; Wilting and Steenbergen, 2016).

In certain conditions, the viral genome can be integrated into the cellular genome. Evidence showed that integration events could happen at a relatively early stage of the neoplasia development, sometimes even before the formation of the low-grade lesions (Peitsaro et al., 2002). Disruption of the E2 ORF frequently occurs in the integrant whereas the E6 and E7 ORFs usually remain intact. Together with the loss of viral episomes, the E2-disrupted integration could cause reduction of regulatory E2 proteins within the cells. Since E2 can negatively regulate expression of E6 and E7 by occupying the binding sites within the long control region, the reduction of E2 will weaken transcriptional repression and contribute to deregulation of the two oncoproteins. As a consequence, the effects induced by E6 and E7 will be augmented. This mechanism is considered to be important for malignant transformation, though about 12.5% of cervical cancers can still derive from cells with exclusively episomal HR-HPV (Doorbar et al., 2015; Xue et al., 2012b).

1.2 HR-HPV-related diseases

Persistent infection with HR-HPV types, if left undetected and untreated, can predispose the host to cancer. Due to the mucosal tropism of the known HR-HPV types, HR-HPV-related cancers frequently occur at anogenital areas and in the mucosa of the mouth and throat. In a recently published report, HPVs have been shown to be linked with nearly all of cervical cancer cases globally. Additionally, HR-HPVs are also linked to 88% of anal cancers, 78% of vaginal cancers, 50% of penile cancers, 35% of oropharyngeal cancers and 25% of vulval cancers (de Martel et al., 2017).

HR-HPV-related cervical disease is the main focus of this project. Cervical cancer ranks as the 4th most common cancer among women worldwide, with incidence rates varying across countries and regions (Bruni et al., 2017). In the UK, thanks to the launch of the National Health Service (NHS) cervical screening program in the late 1980s and the introduction of HPV vaccines in 2008, cervical cancer incidence rates have fallen by almost a quarter since the early 1990s. However, despite the encouraging drop in incidence rates, cervical cancer remains a challenging health care problem in the UK. There are still over 3000 new cervical cancer cases being diagnosed and approximately 900 deaths being reported each year [1]. It is worthwhile to note that more than half of these newly diagnosed cervical patients in the UK are younger than 45 years old and the incidence rates peak in woman of childbearing age. Because of this feature, the economic burden of cervical cancer is large. According to a published report, the estimated average yearly cost for each

1 Statistics about cervical cancer were derived from Cancer Research UK http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/cervical-cancer

cervical cancer patient is about £19,400, including both the personal cost and the costs to the NHS and wider society (Salter and Demos, 2014).

Cervical cancer is manifested by vaginal bleeding, discharge and coital pain. It is normally identified through gynaecological screening tests and diagnosed based on systemic clinical examinations. Pathologically, cervical cancer is classified as squamous cell carcinoma, adenocarcinoma, and other rare cancer types. Squamous cell carcinomas are the most common type and constitute around 70%-80% of all cervical malignancies (Marth et al., 2017). In this project, all the cervical cancer samples tested are squamous cell carcinomas. Clinically, cervical cancer is staged using International Federation of Gynaecology and Obstetrics (FIGO) classification (see Table 1), which is applicable to all pathological types.

As can be seen from the FIGO staging, cervical cancer cells can spread to other parts of the body. Direct invasion of adjacent tissues and organs is the primary route of spread. Lymphatic dissemination occurs less frequently. Haematogenous dissemination is even less common and has only been found in about 5% of cervical cancer patients (Gallup, 2008). For early-stage patients (up to FIGO IIA), surgery is recommended and in certain cases, fertility could be preserved by conisation or simple trachelectomy. In more advanced disease, chemoradiotherapy or palliative chemotherapy is often applied (Marth et al., 2017). As shown in dataset (ICD-10 C53) from ISD Scotland[2], the age-standardised overall relative five-year survival rate for 15-99 year old cervical cancer patients diagnosed in Scotland during 2007-2011 is nearly 60.2%. As the stage goes up, the five-year survival rate drops dramatically.

2 Information Services Division (ISD) Scotland is part of NHS National Services Scotland. Statistics about survival were derived from http://www.isdscotland.org/Health-Topics/Cancer/Cancer-Statistics/Female-Genital-Organ/#cervix

Table 1 The latest FIGO staging for carcinoma of cervix (Pecorelli, 2009)

Stages	Definition
I	The carcinoma is strictly confined to the cervix (extension to the corpus would be disregarded)
IA	Invasive carcinoma which can be diagnosed only by microscopy, with deepest invasion ≤5 mm and largest extension ≤7 mm
IA1	Measured stromal invasion of ≤3.0 mm in depth and extension of ≤7.0 mm
IA2	Measured stromal invasion of >3.0 mm and not >5.0 mm with an extension of not >7.0 mm
IB	Clinically visible lesions limited to the cervix uteri or pre-clinical cancers greater than stage IA
IB1	Clinically visible lesion ≤4.0 cm in greatest dimension
IB2	Clinically visible lesion >4.0 cm in greatest dimension
II	Cervical carcinoma invades beyond the uterus, but not to the pelvic wall or to the lower third of the vagina
IIA	Without parametrial invasion
IIA1	Clinically visible lesion ≤4.0 cm in greatest dimension
IIA2	Clinically visible lesion >4.0 cm in greatest dimension
IIB	With obvious parametrial invasion
III	The tumour extends to the pelvic wall and/or involves lower third of the vagina and/or causes hydronephrosis or non-functioning kidney
IIIA	Tumour involves lower third of the vagina, with no extension to the pelvic wall
IIIB	Extension to the pelvic wall and/or hydronephrosis or non-functioning kidney
IVA	The carcinoma has extended beyond the true pelvis or has involved (biopsy proven) the mucosa of the bladder or rectum. A bullous edema, as such, does not permit a case to be allotted to Stage IV
IVA	Spread of the growth to adjacent organs
IVB	Spread to distant organs

Note: The content of this table was retrieved from
http://pathkids.com/gynae/FIGO%20staging%20revisions%202009.pdf

Benefiting from the effectiveness of the screening tests, nowadays, more and more women at risk of developing cervical cancer are identified at the pre-invasive phase, a relatively long transition state with dysplastic lesions. In order to describe the severity of the cervical dysplasia, the cervical intraepithelial neoplasia (CIN) classification system was introduced by Richart (Reich et al., 2015). According to the histological diagnosis of the samples obtained by biopsy or surgical excision, CIN is classified into three grade levels. Sequentially, CIN grade 1 (CIN1) refers to mild dysplasia that is present only in the lower one third thickness of the stratified epithelium. CIN grade 2 (CIN2) describes moderate dysplasia that affects the lower two thirds of the epithelium. CIN grade 3 (CIN3) defines severe dysplasia that has spread to the entire thickness of the epithelium but has not broken through the basement membrane (Bruni et al., 2017). Since most CIN1 cases could regress naturally without treatment, they are alternatively called low-grade CIN. In contrast, CIN2 and CIN3 are grouped together as high grade CIN due to a much lower rate of spontaneous regression, though a substantial fraction of CIN2 cases do resolve on their own, especially in young women (Castle et al., 2009; Moscicki et al., 2010; Motamedi et al., 2015). In lesions caused by squamous cell abnormalities, the Bethesda system is often applied to report abnormal cytological findings in cytology based screening tests and a two-tiered classification scheme, consisting of low grade squamous intraepithelial lesion (LSIL) and high grade squamous intraepithelial lesion (HSIL), is utilized. In recent years, in order to unify the cytology and histopathology findings and provide a better match for the disease courses, CIN grades have been subsumed under the two-tiered classification scheme. According to the current World Health Organization (WHO) classification for HPV-associated

squamous intraepithelial lesions, CIN1 is subsumed under LSIL (permissive infection with low risk of developing into cancer) while CIN2 and CIN3 are subsumed under HSIL (persistent infection with high risk of progression to cancer). Biomarkers, such as $p16^{INK4a}$, are recommended to be incorporated in the classification of equivocal cases (Reich et al., 2015; Waxman et al., 2012). The WHO classification is also applicable to the precancerous lesion in vulva and vagina. In this thesis, only the two-tiered WHO terminology was used.

1.3 Other factors associated with HR-HPV-related cervical disease

Although HR-HPV infections are quite prevalent in the population, only a small fraction of infected individuals will have persistent infection and even fewer will develop high-grade intraepithelial lesions and cancers (Sheet, 2007). Most HR-HPV infections are transient and are cleared naturally by the immune response without causing any disease. A previous study showed that 70% of newly infected women will be HR-HPV free within one year and as many as 91% will be HR-HPV free within two years (Gerberding, 2004). Despite the fact that persistent infection with HR-HPV types is the leading risk factor for carcinogenesis, persistent infection with HR-HPV types alone is not sufficient to cause cervical cancer. Additional factors, other than HR-HPVs, are required for the cancer to develop and progress. Identifying and tackling these additional factors are expected to accelerate the control and eradication of cervical cancer in conjunction with the prophylactic vaccination and screening programme.

Over the past few decades, many host factors and environmental factors have been proposed and actively investigated as co-factors in HR-HPV-related precancerous cervical disease and cervical cancer. These include smoking, long-term oral contraceptive use, co-infection with other sexually transmitted pathogens (e.g. *C. trachomatis*, *N. gonorrhoeae* and Herpes simplex virus type 2), sexual activities associated with increased HPV acquisition and immunosuppressive conditions (de Abreu et al., 2016; Kapeu et al., 2008; Wang et al., 2009b). Among these factors, genetic variations have long been considered as key explanations for individual differences in disease susceptibility.

According to two nationwide family-based studies, genetic factors accounted for nearly one-fourth of the variation in liability to cervical cancer in the Swedish population (Czene et al., 2002; Magnusson et al., 2000) while the contribution of the shared environment was quite small (Magnusson et al., 2000). A similar proportion (i.e. 24%) was also reported by a recent study based on genome-wide complex trait analysis (Chen et al., 2015). Although genetic factors are not as easy to intervene with as other factors (e.g. infections and behaviours), studies focusing on genetic variations have the potential to enhance our understanding of the disease pathogenesis, to improve triage management of the susceptible population, to identify biomarkers for early detection of cancer, to predict individual responses towards prophylactic vaccinations and treatments, and to facilitate development of novel treatments.

1.4 Single nucleotide polymorphism and HR-HPV-related cervical disease

1.4.1 Single nucleotide polymorphisms

Single nucleotide polymorphism (SNP) is the most common type of genetic variation among humans. SNP can be defined as a variation at a single base pair location in a DNA sequence among individuals (Brookes, 1999). In the human population, the vast majority of identified SNPs are biallelic, which means at each locus, only two of the four DNA nucleotides are observed. For each biallelic SNP, the nucleotide with a frequency above 50% in the studied population is called the major allele while the other is called the minor allele. SNPs with a global minor allele frequency (GMAF) greater than or equal to 5% are considered as common SNPs and have been centrally targeted in previous genomic studies, such as the Phase I and II International HapMap Project. According to the public-domain archive dbSNP run by the NCBI in the USA, more than 10 million common SNPs have been validated in the human genome to date and the number is still increasing (Sherry et al., 2001). In order to unify the description of SNPs, dbSNP assigns an accession ID (rs number), a unique identifier, to each cluster of submitted SNPs mapped to the same position in the genome and routinely performs "rs merge" procedure to reduce the redundancy. Currently, the rs number is the most widely used identifier in the SNP research field (Sherry et al., 2001).

SNPs exist throughout the genome. They can be found in the coding region of the gene, the non-coding region of the gene as well as intergenic regions. SNPs within the coding region of genes can be classified into two groups: synonymous and

non-synonymous SNPs. The latter, which alters the amino acid sequence of the encoded protein, can be further divided into two subtypes (i.e. missense and nonsense SNPs). In general, the genomic distribution of SNPs is not homogenous. They occur more frequently in non-coding regions than coding regions (Barreiro et al., 2008).

SNPs are considered to arise from ancient mutation events and have been retained during long-term evolution as a consequence of natural selection. Most SNPs do not have apparent impact on health or disease and currently have no known functions. However, increasing evidence suggests that some SNPs do act functionally in certain disease processes. They can affect disease susceptibility, disease outcome or treatment response either by modifying gene regulation or by altering the translation of encoded proteins. Although this kind of SNPs is frequently observed in the regulatory region and the coding region of the gene, they can also be found in other parts of the genome. For example, some intergenic SNPs have been shown to be capable of regulating the expression of a nearby gene through spatial associations (Schierding et al., 2014).

The importance of functional SNPs has been shown in many common diseases and their identification has already led to advances in disease diagnosis and treatment. According to a recent keyword search at https://clinicaltrials.gov/[3](an online database listing 251781 clinical trials in 200 countries), there are at least 395 completed or ongoing clinical trials focusing on the clinical applications of disease associated SNPs. Many of these trails have been performed to assess the predictive value of specific SNPs for disease diagnosis and management. For example, SNP profiles

3 The database is run by the U.S. national institutes of health

identifying high-risk individuals have been proposed for breast cancer, venous thrombosis and coronary heart disease (Beaney et al., 2017; de Haan et al., 2012; Evans et al., 2016). SNP markers can now be tested using automated high-throughput methods at relatively low cost. Most advantageously, the SNP profile remains unchanged during a person's lifetime.

Many functional SNPs have also been used to assess drug efficacy and adverse drug responses. Genetic tests for certain polymorphisms have already been implemented into clinical practice. For example, multiple functional SNPs have been identified within the cytochrome P450 2D6 (CYP2D6) gene and shown to affect the activity of the encoded enzyme. Since this enzyme is important for the metabolism of several opioids, changed function often causes adverse effects in patients on opioid therapy. In some cases, especially in children, overproduction of the metabolite morphine could lead to severe respiratory depression and death. Because of these findings, CYP2D6 activity score determined by SNP analysis has now become an essential determinant of opioid therapy. Genetic testing for CYP2D6 polymorphisms before prescribing several types of opioids has been recommended by the Clinical Pharmacogenetics Implementation Consortium in their clinical guidelines (Crews et al., 2014). Similarly, pre-emptive thiopurine methyltransferase genotype tests have been suggested to guide the dosing of thiopurine (Relling et al., 2011). Human leukocyte antigen B genotyping has been recommended prior to the prescription of abacavir (an anti-HIV drug), carbamazepine (an antiepileptic drug) and allopurinol (a gout treatment) (Hershfield et al., 2013; Leckband et al., 2013; Martin et al., 2012).

Inspired by these findings, it is believed that locating and identifying disease-associated functional SNPs will also provide unique opportunities to uncover

the mechanisms of HPV-linked carcinogenesis and aid in the advancement of diagnostics and therapeutics for HR-HPV associated cervical disease.

1.4.2 Genetic association studies

For a given SNP locus, the specific combination of the alleles on the paternal chromosome and maternal chromosome of each individual is termed as the genotype. Genetic association studies have been widely performed to identify disease-associated SNPs through directly comparing the genotype or allele frequencies of certain SNPs between subjects with and without a disease phenotype. According to their design, genetic association studies can be divided into two kinds: family based studies and population based studies. The former is far less popular among the researchers due to the difficulty of sample collection and the issues of statistical power. In this project, only population based studies were used.

Case-control design with unrelated individuals is frequently used in population based genetic association studies. In some cases, extreme phenotype sampling has been employed to increase statistical power. Currently, two main approaches have been used to perform population based genetic association studies, which are the hypothesis-driven candidate gene approach and genome wide association (GWA) studies. Due to advances in genotyping technology, a large number of researchers have shifted their research focus from a candidate gene approach to GWA approaches and some hold the view that the candidate gene approach should be completely replaced by GWA studies. In my opinion, it is improper to simply treat the relationship between candidate gene approach and genomic approach as the

predecessor and the successor, since they are created and utilized for different purposes.

1.4.2.1 GWA studies

I will start with the genomic approach since GWA studies can be conducted without any prior hypothesis or knowledge and screen the entire genome for possible associations between SNPs and the disease being studied. The majority of published GWA studies have been performed using SNP genotyping chips, which are generated based on tagSNPs (see Figure 3 for a schematic diagram of the summary workflow of GWAs).

For a set of SNP loci that are in close proximity, the sequence of alleles on the same chromosome is termed as the haplotype. In our genome, SNPs are not all inherited independently following the Mendelian inheritance. Some alleles tend to occur together more often than the others. As a consequence, the actual frequency of each haplotype usually departs from the expected frequency obtained by simply multiplying the allele frequency at each locus under the assumption of independent random inheritance. The concept of linkage disequilibrium (LD) therefore has been generated to describe the non-random association between SNP alleles at nearby sites in the genome and mathematical equations have been proposed to measure the difference between observed and expected haplotype frequencies (Reich et al., 2001). SNPs in high LD are almost always inherited together. Jointly, these SNPs form a unit called the “LD bin”. Testing one SNP gives information about all other SNPs in the same LD bin so it is not necessary to test all polymorphisms. The tested SNP is the tagSNP for the whole set of SNPs in the same LD bin.

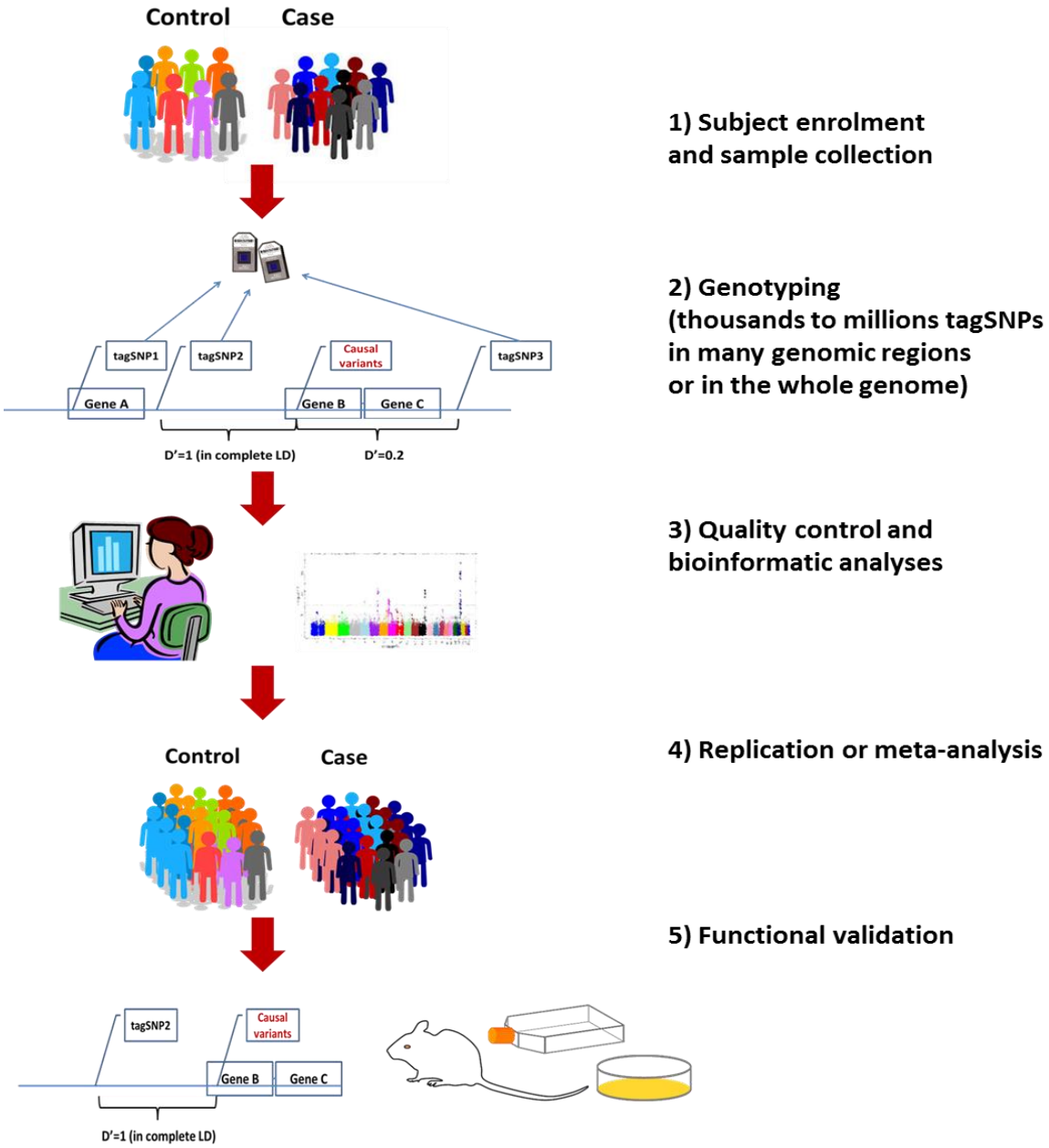


Figure 3 Schematic diagram representing the summary workflow of GWAs

Of note, SNPs in one bin could also be shared by other LD bins. Therefore, the tagSNPs used in commercial SNP chips need to be carefully selected to cover the whole genome in an efficient way. For this reason, tagSNPs are usually not causal SNPs. Once associations have been found between tagSNPs and disease, they cannot be interpreted as direct associations with biological function. Instead the association

signals are more likely due to unknown causal SNPs within the same bin. Additionally, the selection of tagSNPs often relies on data from HapMap projects. As a consequence, the majority of chips can actually only cover common variants or variants with a GMAF above 1% throughout the entire genome and they may have different genomic coverage rate for different ethnic groups due to the various LD patterns among ethnicities.

The above mentioned design strategy and features therefore determine the research questions that can be answered by GWA studies. As expected, a large number of independent susceptible loci for common diseases have been identified over the last decade (Visscher et al., 2017). Many novel common disease-related pathways and genes have been recognized through these findings, which has led to predictive, diagnostic and therapeutic advances. Additionally, based on the identified loci and succeeding complicated summary statistical analyses, we can begin to estimate the genetic architecture[4] for many complex traits, though it still takes time to include the contribution of rare variants and other genetic variations. With an improved understanding of the genetic architecture, we could tailor follow-up studies to further dissect the causal variant and translate the association signal into biological insights.

However, it is important to know that GWA study itself is not designed to identify and dissect the causal variant. For common diseases contributed by numerous variants with low or moderate individual effect size, a substantial increase in the sample size is expected to significantly increase the yield of GWA studies.

4 Genetic architecture of a disease indicates the distribution and the interrelationship of allele frequencies and effect sizes of all causal variants

1.4.2.2 Hypothesis driven candidate gene approach

Candidate gene approaches were widely used before the era of GWA studies in the absence of a clear understanding of the genetic architecture. Unlike GWA studies, the purpose of the candidate gene approach is to identify direct associations caused by plausible functional variants within pre-selected genes and therefore bridge the gap between the genotype to the phenotype (Kwon and Goate, 2000; Zhu and Zhao, 2007). As a hypothesis driven approach, known biological or functional relevance of the pre-selected genes to the disease being studied are necessarily required. Additionally, the tested SNPs within the candidate genes need to be carefully selected. Usually these SNPs have already been confirmed or predicted to be functional so once the associations have been found, it is quite straightforward to establish the genotype-phenotype link and offer biological insights.

The tagSNP strategy has also been applied to the candidate gene approach to cut down costs, especially when multiple potentially functional SNPs in complete LD are present within the candidate genes. There are also situations in which researchers are really interested in certain genes but no known functional or potentially functional SNPs have been identified within these genes. In these situations, normally two methods will be applied. The first is to scan the entire gene using a tagSNP strategy just like a miniature version of a GWA study. The true causal variants can be captured when high density SNP genotyping has been performed, which is usually quite costly. The second is to choose several pleiotropic SNPs (i.e. SNPs that have been shown to be associated with multiple diseases). Since pleiotropy is pervasive, pleiotropic SNPs are more likely to be the causal SNPs or in high LD with the real causal SNPs.

Compared with resource intensive genomic approaches, candidate gene approach based studies are much more affordable and accessible to independent laboratories and researchers with little or no bioinformatics background. Additionally, benefiting from the function centric design, the results obtained from candidate gene based association studies are usually more actionable. However, it is important to know that the power of the candidate gene approach to detect direct associations is highly dependent on the rational selection of candidate genes and SNPs. Currently, for most common diseases, our existing knowledge about aetiological mechanisms and genetic architecture are still not sufficient to provide prefect predictions. Consequently, there is no guarantee that every effort will be rewarded. In addition, many results obtained through candidate gene approaches have been criticized for poor replicability and lack of further validation or follow-on functional analysis. When the sample size of the original study is small or the tested SNPs have not been functionally validated, it is necessary to provide functional evidence and complete the validation steps before making extensive generalizing conclusions.

1.4.3 GWA study findings for HR-HPV-related cervical disease

Over the past 10 years, more than 39000 unique SNP-trait associations have been identified through GWA studies. For many diseases, the findings obtained from GWA studies have refined our understanding of disease pathogenesis and treatment (Visscher et al., 2017). Indeed, the success of GWA study in its first decade is undeniable. However, for researchers interested in HR-HPV-related cervical disease, it is still too early to celebrate.

According to the NHGRI-EBI GWAS Catalog, to date, only 4 susceptible loci have been identified for HR-HPV-related disease, which are 4q12 (rs13117307), 17q21.1 (rs8067378), 6p21.32 (rs4282438, rs9272143, rs73730372, rs3117027 and rs927795) and 6p21.33 (rs2516448) (MacArthur et al., 2016). The first two susceptible loci were identified in a Chinese population while the last one was found in a Swedish population. The only shared susceptible locus between the two populations is 6p21.32. Together these loci can explain only a small fraction of the estimated heritability of cervical cancer (Chen et al., 2015). The details of these loci and subsequent studies derived from these findings are discussed in depth in **Section 3.1**.

The main explanation for the low yields of these GWA studies is their inadequate sample size. A previous study showed that the trait/disease-associated SNPs identified through GWA studies generally have moderate effect size (with a median odds ratio around 1.3). In highly polygenic common diseases, very few SNPs with an odds ratio above 4 have been found (Hindorff et al., 2009). According to a newly published authoritative review for GWA studies, under the assumption of a type I error rate of 5×10^{-8} (conventional genome wide significance threshold) and 80% power, $\sim10^{5}$ samples are required to detect associations caused by SNPs with a MAF<10% and an odds ratio of around 1.2. For SNPs with lower MAF frequency and smaller effect size, the sample size required would be larger (Visscher et al., 2017). Unfortunately, none of the published GWA studies in HR-HPV-related cervical disease have a sample size above 10^{4}, even after pooling the cases from the replication sample set with the initial sample set. It is not surprising at all that the one performed in Japanese population failed to find any SNPs reaching the significance

threshold, since only several hundreds of subjects were genotyped (Miura et al., 2014).

Another issue that has hindered the ability to detect associations in published GWA studies is the heterogeneity of the cases. In the study performed in a Swedish population, the case group includes both carcinoma in situ and invasive cervical cancer cases (Chen et al., 2013). In fact, not limited to one or two GWA studies, similar grouping strategy has been used widely in association studies for cervical cancer. Although this strategy, to a large extent, reduces sampling difficulties, it also substantially decreases the power to detect associations. Additionally, as mentioned previously, the pathogenesis of cervical cancer consists of three transition states: the acquisition of HPV and productive infection, the formation of high-grade cervical intraepithelial lesion and the transition from precancerous lesion to cancer. These transition states are believed to be driven by different underlying mechanisms. Accordingly, the SNPs associated with each transition state could be different. And for a given SNP, its role may differ in different transition states. It is entirely conceivable that associations related to a specific transition state would be missed when cases representing mixed transition states are used.

Although they are not perfect, current GWA studies still provide important information about the genetic architecture and indicate possible directions for the future. Based on summary statistics from GWA study results, Chen and colleagues estimated that 88.9% of the heritability of cervical cancer is tagged by common SNPs (Chen et al., 2015), which implies that we should be able to obtain more new discoveries through GWA studies by using homogeneous samples and by increasing

the sample size to an achievable number. It is only a matter of time, and probably cost.

1.4.4 Findings from candidate gene based association studies

The last two decades have been boom years for candidate gene based association studies. Hundreds of studies have been performed and a large number of SNPs associated with HR-HPV-related cervical disease have been characterised. In recently published studies, SNPs reported to be associated with cervical cancer using candidate gene approaches have been thoroughly reviewed (Chen and Gyllensten, 2015; Martínez-Nava et al., 2016; Mehta et al., 2017; Wang et al., 2015). The majority of disease associated SNPs were found within genes/regions involved in host defence and oncogenesis, such as tumour protein p53 (TP53), cytochrome P450, family 1, subfamily A, polypeptide 1 (CYP1A1), Toll-like receptor 9 (TLR9) , x-ray repair complementing defective repair in Chinese hamster cells 1(XRCC1), methylene tetrahydrofolate reductase (MTHFR), BRCA1-interacting protein 1 (BRIP1) and p21, to name but a few (Chen and Gyllensten, 2015; Martínez-Nava et al., 2016; Wang et al., 2015).

With the rapid development of sequencing technology and bioinformatics tools, large-scale SNP genotyping has also been utilized in candidate gene based association studies. This technique enables researchers to genotype a huge number of SNPs within candidate genes simultaneously in large groups of individuals. As a consequence, some candidate gene based association studies can also be used to indicate new disease related genes based on summary statistics by extending their candidate gene list with genes of no known relevance. For example, Wang and

colleagues evaluated 7140 tagSNPs from 305 candidate genes, which were hypothesized to be involved in DNA repair, viral infection and cell entry, in a population-based cohort study. The study found that six genes (i.e. GTF2H4, DUT, DMC1, OAS3, IFNG and SULF1) and three regions (i.e. TMC6, TMC8 and FLJ35220) were potentially associated with CIN3/cervical cancer. Variants in OAS3, SULF1, DUT, and GTF2H4 were associated with HPV persistence whereas IFNG and TMC6/TMC8 SNPs were associated with progression to CIN3/cancer (Wang et al., 2010). A subsequent study performed by Safaeian and co-workers further evaluated 18310 tagSNPs from 1113 candidate genes hypothesized to be involved in a wide range of cancers. Two genes, RPS19 and PRDX3 were reported to be associated with increased cervical cancer risk and both SNPs (i.e. PRDX3 rs7082598 and RPS19 rs2305809) were associated with CIN3/cancer and the progression (Safaeian et al., 2012). Before the publication of these studies, some of the above genes had never been linked to known pathological processes in HPV infection or cervical cancer.

Though findings from candidate gene association studies vastly outnumber those obtained from GWA studies, they did not lead to clinical advances as effectively as was expected. It is noteworthy that not many of the published associations in candidate gene based association studies have been further validated. According to a recent review, nearly 90% of the genetic associations reported in studies of cervical cancer have been investigated only once (Martínez-Nava et al., 2016). And for those associations that have been investigated in more than one study, only few exhibited consistency among studies and even fewer have been shown to be functionally-relevant. These situations obviously have already deviated from the

original purpose of candidate gene based studies: to bridge the knowledge gap between genetic association and the underlying biological mechanisms.

When comparing the results from published GWA studies and candidate gene based association studies, only a handful of SNPs within the HLA regions of chromosome 6 have been picked out by both approaches. None of the non-HLA SNPs identified by candidate gene association studies have been shown to be significant in current GWA studies (Johanneson et al., 2014). Although a possible explanation for these non-overlapping findings is the low mapping resolution for the current GWA studies, it is unfair to shift all the blame to GWA studies. It is time to raise the standards for reporting associations and recall the original intentions of the candidate gene approach. Especially when irrelevant genes and functionally unknown SNPs are tested, replication or functional validation of the associations should be included as a substantial part of any candidate gene based study.

1.4.5 Now and the future

It is clear that the majority of causal SNPs in HR-HPV-related cervical disease are still unknown. As a promising tool, GWA studies with a substantial increase in the sample size will lead to the discovery of new associations. However, it is hard to predict whether this could be happen in the near future. To reach the goal, stable funding sources and multi-centre and multi-disciplinary collaboration are necessarily required. And the post GWAS analysis will still be quite challenging, especially when the aim is to translate GWAS findings into biological insights. During the waiting time for future GWAS findings, the candidate gene approach will continue to

be a powerful tool for independent laboratories and individual researchers to identify novel targets and establish functional links.

Apart from GWA studies and candidate gene approaches, sequencing based genomic study now becomes the third choice for researchers to identify associations. This approach has a higher mapping resolution than chip based GWA studies and is expected to pick up rare variants missed by GWA studies. Like GWA studies, at the moment, this approach is both expensive and resource intensive. Additionally, due to an even larger amount of data generated, it will be a real challenge to extract association signals from the background noise. In the cervical cancer research field, whole genome sequencing (WGS) has not yet been used to identify associations of SNPs with disease. The Cancer Genome Atlas (TCGA) Research Network has recently published findings from their genomic study based on 228 cervical cancer cases but they mainly focused on somatic genomic alterations, rather than SNPs (Network, 2017). According to what has been done for other common diseases, WGS data could be used as reference panel to impute existing GWAS datasets and consequently boost the number of variants assessed and increase the resolution of the original GWA studies (Luo et al., 2017; Mitt et al., 2017). It is strongly expected that GWAS and WGS jointly will also yield more new associations in HR-HPV-related diseases in the future.

1.5 Chemokines and chemokine polymorphisms in HR-HPV-related cervical disease

1.5.1 Chemokines and chemokine receptors

Chemokines are a group of small chemotactic cytokines that capable of inducing chemotaxis in nearby responsive cells and altering the immune/inflammation responses. To date, about 50 chemokines and 22 cognate chemokine receptors (i.e. 18 typical chemokine receptors and 4 atypical receptors) have been identified (Nagarsheth et al., 2017). According to the location of the first two cysteine residues near the N terminus of the chemokine, chemokines can be divided into four groups: CC group, CXC group, CX_3C group and C group. The CC chemokines contain two adjacent cysteine residues while the CXC chemokines have a non-conserved amino acid between the two cysteines. CX_3CL1 is the only member in the CX3C group and is characterized by having three intervening amino acids between the two cysteines. The C group has two members and they only have one N terminal cysteine. According to their functions, chemokines can also be loosely classified into two types: homeostatic and inflammatory chemokines. Homeostatic chemokines are constitutively expressed in physiological conditions and participate in basal leukocyte trafficking while inflammatory chemokines are predominantly produced in pathological conditions and attract leukocyte towards the site of inflammation. The majority of chemokines can bind to more than one chemokine receptor and vice versa. Together, chemokine and chemokine receptor pairs are involved in a wide variety of processes in infectious diseases and cancers (Cardona et al., 2013).

1.5.2 Chemokine and chemokine receptors in HR-HPV-related cervical disease

The roles of cytokines have been extensively investigated in HR-HPV-related cervical disease (Paradkar et al., 2014). However, unlike other cytokines, the role of chemokines in HR-HPV cervical cancer has rarely been investigated. Previous work done in our group showed that many chemokines were up-regulated in HR-HPV induced high-grade dyskaryosis. Particularly, six chemokines (i.e. CXCL1, CXCL8, CXCL10, CXCL12, CCL2 and CCL3) were found to be up-regulated in HR-HPV positive high grade cervical disease on both the Luminex® and the proteomic array platforms (unpublished data).

CXCL1 (or called growth-regulated oncogene alpha, GROα) is a potent pro-angiogenic chemokine. Sales et al found that CXCL1 released by HeLa cells in response to seminal plasma was capable of regulating vascular function in *in vitro* angiogenesis assays (Sales et al., 2012). Serum levels of CXCL1 have been reported to be up-regulated as disease progress in patients with cervical cancer (Zhang et al., 2014). Contrary to CXCL1, CXCL10 (also called interferon-gamma-inducible protein 10, IP 10) is an anti-angiogenic chemokine. Yang et al showed that CXCL10 could down-regulate microvessel density and increase apoptotic rate in an *in vivo* mouse model of cervical cancer (Yang et al., 2009). In cultured condyloma acuminatum tissues, Li et al found that CXCL10 could induce apoptosis and reduce HPV expression (Li et al., 2009). Furthermore, CXCL10 levels were found to be inversely correlated with vascular endothelial growth factor (VEGF) levels in patients with cervical cancer (Sato et al., 2007). CXCL8 (also called interleukin 8)

plays key role in many physiological and pathological processes. It is of particular interest in this project. Details of this chemokine and its receptor CXCR1 and CXCR2 are discussed in depth in **Section 5.1**. CCL2 (also called monocyte chemotactic protein 1, MCP1) is a pro-inflammatory chemokine, which is able to recruit monocytes, memory T cells and dendritic cells to the site of inflammation. Elevated serum CCL2 has been associated with advanced cervical cancer and lymphatic metastasis (Lebrecht et al., 2001). The absence of CCL2 expression in cervical cancer has been shown to be beneficial (Zijlmans et al., 2006). In a novel molecular cascade identified by Schroer et al in cervical cancer, activation of signal transducer and activator of transcription 3 (STAT3) led to a strong up-regulation of CCL2, which amplified matrix metallopeptidase 9 (MMP-9) expression and contributed to the switch from precancerous lesions to malignancy (Schroer et al., 2011). CXCL12 (also called, stromal cell-derived factor 1, SDF-1) is a homeostatic chemokine, which is strongly chemotactic for lymphocytes and highly associated with tumorigenesis (Kryczek et al., 2007). The over-expression of CXCL12 and its receptor CCR4 has been correlated with tumour progression in cervical cancer (Huang et al., 2013; Wei et al., 2007). In general, the mechanisms by which chemokines influence HPV clearance and carcinogenesis are still far from being fully elucidated.

1.5.3 Chemokine polymorphisms in HR-HPV-related cervical disease

The essential role of chemokine in immune responses and other multiple physio-pathological processes makes them ideal candidate genes for genetic association studies. Conversely, locating and identifying functional SNPs in these

chemokine genes would help to explain how these genes influence HPV clearance and carcinogenesis. However, there are very few reports on variations in chemokine genes or chemokine receptor genes in HR-HPV-related cervical diseases. To the best of my knowledge, only CCL2 rs1024611, CCR2 rs1799864, CXCL8 rs4703, CCR5 rs333 (a frame-shift variant) and several variants in CXCL12 have been linked to HR-HPV-related cervical diseases to date (Catarinoa et al., 2007; Chatterjee et al., 2010; Coelho et al., 2005; Ivansson et al., 2007; Maley et al., 2009; Singh et al., 2008; Wu et al., 2013a; Wu et al., 2011; Zheng et al., 2006) , which seems incommensurate with the boom in studies on genetic polymorphisms. Furthermore, for some of the SNPs mentioned above, there is little consensus among existing studies. For example, the impacts of CCR2 rs1799864 on the pathogenesis of cervical cancer are still not clear. Coelho et al and Ivansson et al reported that the CCR2-64I variant (corresponds to the A allele) played a protective role in the development of cervical cancer (Coelho et al., 2005; Ivansson et al., 2007), whereas Chatterjee et al suggested A allele carriers were at a higher risk of cervical cancer (Chatterjee et al., 2010),whilst according to the studies conducted by Singh et al and Zheng et al, the CCR2-64I variant showed no association with either HPV infection or cervical cancer development (Singh et al., 2008; Zheng et al., 2006). The findings are contradictory but this might be explained by inadequate sample sizes, ethnic and geographic variation. Therefore, studies concerning the role of SNPs in chemokine genes or chemokine receptor genes in HR-HPV-related cervical diseases are needed. However, as mentioned previously, caution must be taken in regards to confounding variables before reporting any association.

1.6 Study hypothesis, aim and objectives

This PhD project was instigated by data from published genetic association studies and from the preliminary chemokine profile study in the laboratory of one of my supervisors, Prof. Sarah Howie.

The overarching hypothesis of the project is:

There are functional SNPs within genes involved in host defence and/or oncogenesis that are associated with HR-HPV-related cervical cancer and precancerous cervical disease.

The aim of the project is to identify functional variants in genes/regions of confirmed associations and to investigate the mechanisms by which they may contribute to HR-HPV-related cervical disease

The objectives for the project:

1. Select a proper panel of candidate SNPs for a small scale candidate gene based association study through literature mining.

2. Investigate the associations between selected SNPs with HR-HPV-related cervical disease through a custom oligo-microarray.

3. Validate the genetic associations found in the SNP array in the same set of samples and additional samples using other genotyping methods.

4. Design functional assays to investigate the role of the significant SNPs identified from the SNP array in HR-HPV-related cervical disease

Chapter 2
Materials and Methods

2.1 Ethics statement

All Samples used in this study were anonymised and routinely taken. Samples used in the genetic association study were extracted genomic DNA (gDNA) (where available), liquid-based cytology (LBC) cervical samples or formalin-fixed, paraffin-embedded (FFPE) sections from biopsies provided by Scottish HPV archive and Lothian NRS Bioresource. The provision of extracted DNA samples and subsequent research was conducted under the approval held by the NHS Lothian NRS Bioresource (East of Scotland REC Ref 15/ES/0094).

2.2 Materials

2.2.1 Clinical samples

A panel of routinely taken clinical samples were tested in the SNP array, which incorporated 475 cytologically-normal and HR-HPV negative cases, 413 HR-HPV positive HSIL cases and 62 HR-HPV positive cervical squamous cell carcinomas (see **Section 2.3.2.1** for sample size estimation). The 475 controls were gDNA samples obtained directly from Lothian Year 2000 surveillance collection. These samples were previously extracted from LBC samples using either the QIAamp DNA Mini Kit (Cat#:51306, QIAGEN Ltd., Manchester, UK) or crude DNA extraction for unrelated reasons. All the HSIL cases were obtained from HPV Surveillance collection, of which 264 were extracted by me from LBC samples using MasterPure™ Complete DNA and RNA Purification Kit. The rest of HSIL cases were gDNA extracted from FFPE samples. These samples were also previously extracted from FFPE samples using QIAamp DNA FFPE Tissue Kit (Cat#:56404, QIAGEN Ltd., Manchester, UK) for unrelated reasons. Cervical cancer samples were

all obtained from the Scottish Invasive Cervical Cancers collection and extracted by me using GeneRead DNA FFPE Kit. These samples were provided with relevant anonymised patient information, including the age at diagnosis, the date on which the specimen was received, the date on which the genomic DNA was extracted (when applicable), HR-HPV types present[5] and histologic disease grade (see Table 2 for details). For the cancers, International Federation of Gynaecology and Obstetrics (FIGO) stage, TNM category, vital status, survival time after the biopsy were taken, date of death and the cause of death were also provided.

Table 2 Frequency distribution of select characteristics in HR-HPV high-grade cases and HR-HPV negative, cytologically-normal controls

Variables	Cases (n=475)	Controls (n=475)
Age, yrs (mean ± SD)	23.88 ±11.57	38.18±11.43
Source, n (%)		
FFPE	211(44.4%)	0(0%)
LBC samples	264(55.6%)	475(100%)
HR-HPV, n (%) [6]		
HPV16+	336(70.7%)	
HPV16-	139(29.3%)	
Histologic gradations, n (%)		
CIN2+	382(80.4%)	
CIN3	31(6.5%)	
Squamous carcinoma	62(13.1%)	

In order to investigate the role of significant SNPs in a broader spectrum of HPV-related diseases, genotyping results of additional clinical samples were included in the statistical analysis of the validation step, including 83 HPV positive, cytologically normal LBC samples (average age: 20), 21 HPV positive cervical

5 HR-HPV types, including type 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 and 68, are of most concern in this project.

6 HPV16 is the biggest causal agent of cervical cancer and is the most widely investigated type.

cancer samples (average age: 45.2±12.3), 129 HPV positive vulval intraepithelial neoplasia (VIN) samples (average age: 47.5±12.4) and 23 HPV positive vulval cancer samples (average age: 56.6±12.2). These additional samples were genotyped by Luhan Jiang and Dr Michelle Etherson respectively as part of their Master projects in the HPV group. The results were carefully re-inspected by me before pooling with the array results. The use of these data was conducted under the approval held by the NHS Lothian NRS Bioresource (East of Scotland REC Ref 15/ES/0094) and National Health Service Greater Glasgow and Clyde (NHSGGC) Bio-repository (NHSGGC Bio-repository Ref number 96).

2.2.2 Reagents and Chemicals

All chemicals came from Sigma-Aldrich, Poole, UK, unless otherwise stated. All the primer/probe sets for the genotyping PCR were purchased from TIB Molbiol, Berlin, Germany. GAPDH Primer & Probe set (Cat#: HK-DD-hu-900) was obtained from PrimerDesign Ltd, Chandler's Ford, UK.

Reagents	Manufacturer
Agarose powder	Cat#:BIO-41025, Bioline, London, UK
SYBR®safe DNA gel stain	Cat#:S33102, Invitrogen, Paisley, UK
GeneRuler 1 kb DNA Ladder	Cat#:SM0312, ThermoFisher Scientific, Paisley, UK
Dulbecco's Modified Eagle's medium(DMEM) 1X	Cat#:21969-035, ThermoFisher Scientific, Paisley, UK
GIBCO® Dulbecco's Phosphate-Buffered Saline (DPBS) 1X	Cat#:4190-094, ThermoFisher Scientific, Paisley, UK
Gibco® RPMI Medium 1640 1X	Cat#:31870-025, ThermoFisher Scientific, Paisley, UK
Fetal Bovine Serum, Qualified, Heat Inactivated	Cat#:10500-064, ThermoFisher Scientific, Paisley, UK
0.05% Trypsin - EDTA (1x) + phenol red	Cat#:25300-054, ThermoFisher Scientific, Paisley, UK
L-Glutamine	Cat#:25030-024,

	ThermoFisher Scientific, Paisley, UK
Pen-Step	Cat#:15140-122, ThermoFisher Scientific, Paisley, UK
G418 (solution)	Cat#:ant-gn-1, InvivoGen Europe, Toulouse, France
Blasticidin (solution)	Cat#:ant-bl-1, InvivoGen Europe, Toulouse, France
APC Anti-human CD181 (CXCR1)	BioLegend, London, UK
PE anti-human CD182 (CXCR2)	BioLegend, London, UK
Lipofectamine 2000	Cat#:11668027, ThermoFisher Scientific, Paisley, UK
ProLong® Gold Antifade Mountant with DAPI	Cat#:P36931, ThermoFisher Scientific, Paisley, UK
EDTA solution	Cat#:E8008, Sigma-Aldrich Company Ltd., Gillingham, UK
Human TruStain FcX™	Cat#:422301,Biolegend, London, UK
Recombinant human IL-8	Cat#:208-IL-010/CF, R&D Systems, Abingdon, UK
Opti-MEM I Reduced Serum Media	Cat#:31985062, ThermoFisher Scientific, Paisley, UK

2.2.3 Media, buffers and solutions

LB medium (Cat#:L3522-1KG, Sigma-Aldrich Company Ltd., Gillingham, UK):10g/L Tryptone, 10g/L NaCl and 5g/L Yeast Extract

LB Broth with agar (Cat#:L3147-1KG, Sigma-Aldrich Company Ltd., Gillingham, UK): 15g/L Agar, 10g/L Tryptone,10g/L NaCl and 5g/L Yeast Extract

2x YT medium (Cat#:Y2377-250G, Sigma-Aldrich Company Ltd., Gillingham, UK):16g/L Tryptone, 10 g/L Yeast Extract and 5g/L NaCl

1x TAE buffer: 40 mM Tris, 20 mM acetic acid and 1 mM EDTA.

Alkaline lysis Solution I: 50 mM glucose, 25 mM tris-HCl (pH 8.0) and 10 mM EDTA (pH 8.0)

Alkaline lysis Solution II: 0.2 N NaOH and 1% (w/v) SDS

Alkaline lysis Solution III: 5 M potassium acetate (60.0 ml), glacial acetic acid (11.5 ml) and H2O (28.5 ml)

2.2.4 Kits and enzymes

Kits

Kits	Manufacturer
MasterPure™ Complete DNA and RNA Purification Kit	Cat#:MC85200, Cambio, Cambridge, UK
GeneRead DNA FFPE Kit	Cat#:180134, QIAGEN Ltd., Manchester, UK
RNeasy Micro Kit	Cat#:74004, QIAGEN Ltd., Manchester, UK
KOD Hot Start DNA Polymerase kit	Cat#:71086-3, Merck Millipore, Watford, UK
QIAprep Spin Miniprep Kit	Cat#:27104, QIAGEN Ltd., Manchester, UK
LightCycler® FastStart DNA Master HybProbe kit	Cat#:12239272001, Roche, Lewes, UK
QuantiTect Reverse Transcription Kit	Cat#:205311, QIAGEN Ltd., Manchester, UK
PureYield™ Plasmid Midiprep Kit	Cat#:A2492, Promega UK, Southampton, UK
Cell Counting Kit-8 (500tests)	Cat#:CK04-05 Munich, Germany

Enzymes:

XbaI	Cat#:R0145S, New England Biolabs, Hitchin, Herts, UK
NheI	Cat#:R0131S, New England Biolabs, Hitchin, Herts, UK
DpnI	Cat#:R0176S, New England Biolabs, Hitchin, Herts, UK
EcoRV	Cat#:R0195S, New England Biolabs, Hitchin, Herts, UK
Gateway® LR clonase II enzyme mixture	Cat#:11791020, ThermoFisher Scientific, Paisley, UK

2.2.5 Biological materials

Plasmids

All the plasmids used in this project were provided by Professor Jürgen Haas, Division of Infection and Pathway Medicine, University of Edinburgh, UK.

Bacteria

All the bacteria strains used in this project were provided by Professor Jürgen Haas, Division of Infection and Pathway Medicine, University of Edinburgh, UK.

Cell lines

Cervical cancer cell lines: HeLa (HPV18 positive) and SiHa (HPV16 positive) cell lines were obtained from ATCC (LGC standards, Teddington, UK); the CaSki (HPV16 positive) cell line was obtained from Dr Kate Cuschieri, Scottish HPV Reference Laboratory, Edinburgh, UK.

Other cell lines: the human monocytic cell line Mono Mac 6 (MM6) was obtained from DSMZ, Braunschweig Germany; human embryonic kidney cell lines HEK293 and 293T were provided by Professor Jürgen Haas, Division of Infection and Pathway Medicine, University of Edinburgh, UK.

All cell lines were tested and shown to be mycoplasma free.

2.2.6 Equipment and consumables

Equipment	Manufacturer
LightCycler® Nano Instrument	Roche, Lewes, UK
Milli-Q® Academic Water Purification system	Merck Millipore, Watford, UK
Mr. Frosty™ Freezing Container	ThermoFisher Scientific, Paisley, UK
FumeGuard Controller	FUMAIR, East Kilbride, UK
GeneAmp® PCR System 9700	Applied Biosystems, Warrington, UK
Eppendorf 5415D centrifuge	Eppendorf UK Ltd., Stevenage, UK
Eppendorf Thermomixer® R Dry Block Heating and Cooling Shaker	Eppendorf UK Ltd., Stevenage, UK
Multitron Standard incubation shaker	Infors HT, Bottmingen, Switzerland
Raven incubator	LTE Scientific, Greenfield, Oldham, UK
Eppendorf 5810R centrifuge	Eppendorf UK Ltd., Stevenage, UK
M330 UV-Visible Spectrophotometer	Camspec Limited, Sawston, Cambridge, UK
Sigma 6K15 high speed centrifuge with 12256-H angle rotor	DJB Labcare Ltd, Newport Pagnell, UK
Eppendorf Eporator®	Eppendorf UK Ltd., Stevenage, UK
U:GENIUS3 UV transilluminator	Syngene, Cambridge, UK
NanoDrop® ND-1000 UV-Vis Spectrophotometer	Nanodrop Technologies, Ringmer, UK
Applied Biosystems® StepOne™ Real-Time PCR System	ThermoFisher Scientific, Paisley, UK
BioMAT® 2 Safety Cabinet	Contained Air Solutions Ltd, Manchester, UK
Galaxy R+ incubator	RS Biotech Laboratory Equipment Ltd., Irvine, UK
QuantStudio 12K Flex Real Time PCR System	Applied Biosystems, Cheshire, UK
Leica TCS SP5 confocal microscope	Leica Microsystems Ltd, Milton Keynes, UK
EVOS™ digital inverted microscope	Advanced Microscopy Group, Bothell, WA
LSR Fortessa analyser	BD Biosciences, Oxford, UK
FACSAria II sorter	BD Biosciences, Oxford, UK
Waterbath	Grant Instruments, Shepreth, UK

Consumables	Manufacturer
Corning® Cell culture flask (T25,T75 and T162cm^2)	Sigma-Aldrich, Poole, UK
Corning® Costar® TC-Treated Multiple Well Plates (Flat bottom,6,12,24,48 and 96-well)	Sigma-Aldrich, Poole, UK
Falcon tube (15,50ml)	Corning Ltd, Lakeside, UK
Corning® tissue-culture treated culture dishes (D × H 100 mm × 20 mm)	Sigma-Aldrich, Poole, UK
Sterilin™ Standard 90mm Petri Dishes	ThermoFisher Scientific, Paisley, UK
Corning® Costar® Stripette® serologic al pipettes (1,5,10,25 and 50ml)	Sigma-Aldrich, Poole, UK
Tips (Non-sterile,10,20,200 and 1000µl)	Gilson U.K. Dunstable, UK
Tips (Sterile,10,20,200 and 1000µl)	Greiner Bio-One Ltd, Stonehouse, UK
Microcentrifuge Tubes(0.5,1.5 and 2ml)	Eppendorf UK Limited, Stevenage,UK
Gene Pulser®/MicroPulser™ electroporation cuvette	Cat#:1652082, Bio-Rad Laboratories Ltd., Watford, UK
96-well deep well plate	VWR International Ltd
low profile 96 well PCR plates	Cat#:AB0600, ThermoFisher Scientific, Paisley, UK
Nalgene™ PPCO Centrifuge Bottles with Sealing Closure	Cat#:3141-0250 ThermoFisher Scientific, Paisley, UK
MicroAmp® Fast Optical 48-Well Reaction Plate	Cat#:4375816 ThermoFisher Scientific, Paisley, UK
MicroAmp® 48-Well Optical Adhesive Film	Cat#:4375323 ThermoFisher Scientific, Paisley, UK
Low profile PCR tubes, strips of 8	Cat#:AB0771, ThermoFisher Scientific, Paisley, UK
Low Profile Tubes and Flat Caps, strips of 8	Cat#:AB0773, ThermoFisher Scientific, Paisley, UK
8-well chamber slide	Cat#: PEZGS0816, Merck Millipore, Hertfordshire, UK
Corning™ Falcon™ Round-Bottom Polystyrene Tubes 5ml	Cat#:352052, Fisher Scientific UK Ltd, Loughborough, UK
Hanging inserts (8.0µm pore, 6.5mm diameter)	Cat#: MCEP24H48, Merck Millipore, Hertfordshire, UK
Falcon™ Cell Strainers, 40µm Nylon	Fisher Scientific Ltd, Loughborough, UK
Rectangular cover glasses, Menzel Gläser 24×50 mm	Cat#:P10143263NR1 ThermoFisher Scientific, Paisley, UK
Gloves, NITRILE, S	Cat#:112-2371,VWR International Ltd
Fisherbrand™ Disposable Pipette Basins	Cat#:13-681-500, Fisher Scientific UK Ltd, Loughborough, UK
FastRead 102™ Disposable plastic counting slides	Immune Systems Ltd, Paignton, UK
Nunc® CryoTubes® 1.8ml	Cat#:V7634, Sigma-Aldrich, Poole, UK

2.3 Methods

2.3. 1 General methods

2.3.1.1 Preparation of chemically competent *Escherichia coli DH10β*

10µl glycerol stock of the *Escherichia coli DH10β* was streaked onto an antibiotic-free lysogeny broth (LB) agar plate and cultured overnight at 37°C in a Raven incubator. Single colonies were picked and cultured in 10ml LB medium at 200rpm, 37°C in a Multitron Standard incubation shaker. The overnight culture was inoculated in 2x YT medium (1:200) and the bacterial density was closely monitored by measuring optical density (OD) at a wavelength of 600 nm using Camspec M330 UV-Visible Spectrophotometer. Once the OD_{600} reached 0.5, bacterial growth was halted by swirling the flask vigorously on ice for 15min. *DH10β* bacteria were pelleted by centrifuging 10min at 4000rpm, 4°C using Sigma 6K15 high speed centrifuge with 12256-H angle rotor and resuspended in 20ml cold sterile $CaCl_2$ (100mM). After incubation on ice for 30min, cells were pelleted again by centrifuging 10min at 4000rpm, 4°C, resuspended at a concentration of 2×10^9 cells/50µl in sterile $CaCl_2$ (100mM) with 15% glycerol and aliquoted in 50µl volumes in a dry ice-ethanol bath. The competency of the bacteria was checked by transforming one aliquot of the cells with 1pg/µl plasmid vector as described in Section 2.3.1.3. Cells with high transformation efficiency (~10^9cfu/µg) were used for the following experiments.

2.3.1.2 Preparation of electrocompetent *Escherichia coli Stbl3*

Glycerol stock of the *Escherichia coli Stbl3* was streaked onto an antibiotic-free LB agar plate and cultured overnight at 37℃ as described in **Section 2.3.1.1**. The overnight culture was inoculated in 2x YT medium (1:200) and the bacterial density was closely monitored by measuring OD at a wavelength of 600 nm. Once the OD_{600} reached 0.5, bacterial growth was halted by swirling the flask vigorously on ice for 10min. *Stbl3* bacteria were pelleted by centrifuging 10min at 6000rpm, 4℃ using Sigma 6K15 centrifuge with 12256-H angle rotor and washed twice with cold sterile H_2O by centrifuging by centrifuging 10min at 6000rpm, 4°C. The bacteria were then resuspended at a concentration of 5×10^9 cells/100µl in 10% Glycerol and aliquoted in 100µl volumes in a dry ice-ethanol bath. The competency of the bacteria was checked by transforming with 1pg/µl plasmid vector as described in **Section 2.3.1.4**. Bacteria with high transformation efficiency (~10^{11}cfu/µg) were used for the following experiments.

2.3.1.3 Transformation of chemically competent *Escherichia coli DH10β*

50 µl of chemically competent *DH10β* bacteria were taken out of -80℃ and thawed on ice. Once thawed completely, 1-2µl of plasmid DNA was added and incubated on ice for 30min, followed by heat shock at 42℃ for 45sec and incubation on ice for 2min. Then the mixture was recovered with 500µl antibiotic-free LB medium and incubated at 37℃ in an Eppendorf Thermomixer® R Dry Block Heating and Cooling Shaker at 800rpm for an hour. After incubation, the mixture was plated directly onto a 10cm LB agar plate containing the appropriate antibiotic. The plate was then cultured in a Raven incubator at 37℃ overnight. Colonies were picked the

next day in appropriate volume of LB medium with the selection antibiotic (1.2ml/colony for alkaline lysis and 5ml/colony for QIAprep Spin Miniprep).

2.3.1.4 Transformation of electrocompetent *Escherichia coli Stbl3*

100µl of electrocompetent *Stbl3* were taken out of -80℃ and thawed on ice. Once thawed completely, 2µl of plasmid DNA was added and the mixture was transferred to an ice-cold Gene Pulser®/MicroPulser™ electroporation cuvette. The mixture was then pulsed in Eppendorf Eporator® at 1.8kV and recovered immediately in pre-warmed antibiotic-free LB medium. After incubation at 37℃ in the Eppendorf Thermomixer at 800rpm for an hour, the mixture was plated and cultured as described in Section 2.3.1.3.

2.3.1.5 Glycerol stock preparation and storage

For lab stock maintenance, glycerol stock was always prepared before DNA extraction by mixing 300µl liquid overnight bacterial culture with same volume of 50% glycerol. Once made, each glycerol stock was labelled with plasmid name and date, and then immediately transferred to a -80℃ freezer for long term storage.

2.3.1.6 Plasmid DNA extraction

2.3.1.6.1 Small scale plasmid DNA extraction

Single bacterial colonies were inoculated into 5ml LB medium with appropriate antibiotic within a 15ml Falcon conical tube and cultured overnight at 37℃ , 160rpm in the Multitron Standard incubation shaker. Next day, glycerol stock was made for each sample as described in Section 2.3.1.5 for long term storage. The rest of the

culture was centrifuged at 3100G for 5min to pellet the bacteria. Plasmids were then extracted using QIAprep Spin Miniprep Kit following the manufacturer's instructions

2.3.1.6.2 Small scale plasmid DNA extraction using alkaline lysis method

Single bacterial colonies were inoculated into 1.2ml LB medium with appropriate antibiotic within 96 well deep-well plate and cultured overnight at 37℃ , 160rpm in the Multitron Standard incubation shaker. Next day, 200µl of culture from each well was transferred to a 96 well plate and stored at 4℃ as a short term stock. The rest of the culture was centrifuged at 2200G for 10min to pellet the bacteria. 300µl of alkaline lysis solution I was added to each well to resuspend the bacteria, followed by 300µl of alkaline lysis solution II for bacterial lysis and 300µl ice-cold alkaline lysis solution III for neutralization (see **Section 2.2.3** for details). Supernatant was then collected by centrifuging at 2200G for 30min and mixed with 600µl of isopropanol overnight at -20ºC to precipitate the plasmid DNA. On day 3, a DNA pellet was obtained by centrifuging at 2200G for 45min, rinsing with 500µl 70% ethanol and air-drying. Finally, the pellet was eluted in 35µl ddH_2O.

2.3.1.6.3 Large scale plasmid DNA extraction

A small volume of a glycerol stock (normally 15-20µl) was inoculated into 200ml LB medium with appropriate antibiotic within an Erlenmeyer flask and cultured overnight at 37℃ , 200rpm in the Multitron Standard incubation shaker. Next day, glycerol stock was made for each sample as described in **Section 2.3.1.5** for long term storage. The rest of the culture was poured into a Nalgene™ PPCO Centrifuge bottle (250ml) and centrifuged at 3100G for 5min to pellet the bacteria. Plasmids

were then extracted using PureYield™ Plasmid Midiprep Kit following the manufacturer's instructions.

2.3.1.7 Agarose gel electrophoresis

Agarose gel electrophoresis was performed to separate DNA fragments generated by diagnostic restriction enzyme digestion. 1% Agarose gel shows good resolution for linear DNA ranging from 500-10000bp and therefore was chosen for this project. The 1% agarose gel was made by mixing 1.7g agarose powder with 170ml homemade TAE (1x) buffer in a microwavable flask and heating in a microwave oven at full power until the agarose powder was completely dissolved. For visualization of DNA fragments, 4µl SYBR®safe DNA gel stain was added after cooling down the gel solution to about 50℃ . Then, the gel solution was poured into a gel tray with a well comb in place. Once completely solidified, the gel was transferred to the electrophoresis tank filled with TAE (1x) buffer. After removing the comb, 7µl GeneRuler 1kb DNA Ladder along with each set of DNA digests were loaded into wells and run at 100V for 50-70min. Picture of the gel was taken under the UV light of the U:GENIUS3 UV transilluminator to visualize the DNA fragments.

2.3.1.8 Cell culture

2.3.1.8.1 Resuscitation of frozen cell lines

To retrieve cells for culture, frozen cells were quickly thawed in a 37℃ waterbath and pelleted by centrifuging at 190G for 5min. Then, the cells were washed with

PBS at 190G for 5min, resuspended in complete growth medium and transferred to culture flasks.

2.3.1.8.2 Cell maintenance

HeLa and MM6 cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 growth medium plus 10% v/v fetal bovine serum (FBS), 1% L-glutamine, 100µg/ml streptomycin and 100U/ml penicillin (37° C, 5% v/v CO_2, 100% humidity in a Galaxy R+ incubator). HEK293T cells were maintained in DMEM growth medium plus 10% FBS, 1% L-glutamine, 100µg/ml streptomycin, 100U/ml penicillin and 500µg/ml G418. Other cells were all maintained in DMEM plus 10% FBS, 1% L-glutamine, 100µg/ml streptomycin and 100U/ml penicillin. Medium was normally changed every two days. Once reaching 80%-90% confluence, adherent cells were detached with Trypsin-EDTA solution, washed with PBS by centrifuging at 190G for 5min, resuspended in fresh growth medium and transferred into new flasks with at a ratio of 1:5 or seeded as required for experiments. Non-adherent cells were split by spinning down at 190G for 5min, washed with PBS and resuspended in fresh growth medium with a ratio of 1:8.

2.3.1.8.3 Cell counting

Cell numbers were determined using FastRead 102 disposable counting chambers following the manufacturer's instructions. Briefly, 9µl aliquoted single cell suspension was added into a chamber of the slide. Cell numbers within at least four complete 4×4 grids were counted under a light microscope at 10x magnification.

Concentration of the cell suspension was calculated as: average cell counts in a 4×4 grid $\times 10^4 \times$sample dilution (if any)/ml.

2.3.1.8.4 Cryopreservation of cell lines

Cells were detached from culture flasks and counted as described in **Section 2.3.1.8.3**. Then, cells were resuspended at a concentration of 1 x 10^7 cells/ml in ice-cold FBS plus 10% Dimethyl sulfoxide (DMSO). Cell suspension was aliquoted in 1.5ml volumes in Nunc® CryoTubes and gradually frozen within Mr. Frosty™ Freezing Container at -80°C. Once completely frozen, cells were transferred to a liquid nitrogen tank for long-term storage.

2.3.1.9 DNA extraction

Genomic DNA was extracted from Cell samples (including LBC samples) and FFPE samples using MasterPure™ Complete DNA and RNA Purification Kit and GeneRead DNA FFPE Kit respectively following the manufacturer's instructions. Purity and concentration of each genomic DNA sample was quantified using a NanoDrop® ND-1000 UV-Vis Spectrophotometer based on the absorbance at 260nm and 280nm. Sample purity was assessed according to the ratio of absorbance at 260 nm and 280 nm, which was normally near 1.8. Sample concentration was automatically calculated based on the absorbance at 260 nm and displayed in ng/μl.

2.3.1.10 RNA extraction and cDNA synthesis

Ribonucleic acid (RNA) was extracted from cell samples using the RNeasy Micro Kit following the manufacturer's instructions and quantified using a NanoDrop® ND-1000 UV-Vis Spectrophotometer. Sample purity was assessed according to the

ratio of absorbance at 260 nm and 280 nm, which was normally near 2.0. Sample concentration was automatically calculated based on the absorbance at 260 nm and displayed in ng/µl. After determination of the concentration, cDNA was synthesized from 500ng RNA using QuantiTect Reverse Transcription Kit following the manufacturer's instructions.

2.3.1.11 Quantitative PCR

Quantitative PCR (qPCR) reactions were performed in MicroAmp® Fast Optical 48-Well Reaction Plate on the Applied Biosystems® StepOne™ Real-Time PCR System (See Table 3 and 4 for the detailed reaction setup and thermal profile). Data acquisition and post-acquisition analysis was performed using StepOne™ Software v2.3 (ThermoFisher Scientific, Paisley, UK).

Table 3 Standard qPCR reaction setup

PCR Component	Volume
TaqMan® Universal PCR Master Mix	*6.5 µl*
PCR-grade H_2O	*5.5µl*
Prime& probe Set	*1µl*
cDNA template	*2 µl*
Total reaction volume	*15µl*

Table 4 Standard thermal profile for qPCR

Step	Thermal profile
1.Polymerase activation	*95°C for 10 min*
2.Denature	*95°C for 15 sec*
3.Anneal/extend	*60°C for 60 sec*
5.Repeat steps 2–4	*39 times (40 in total)*

2.3.2 SNP genotyping array

2.3.2.1 Sample size estimation for the SNP array

The estimated global minor allele frequencies (GMAF) of our candidate SNPs mainly range from 0.05 to 0.4. Sample size for the SNP array was estimated using Quanto version 1.2.4 under the assumption of 0.01% disease prevalence, 1:1 case/control ratio, a range of odds ratio between 1.5 to 3 (RG), 5% Type I error rate and power at 80% under the log-additive genetic model (see Appendix 1 for the details). The log-additive model is the most frequently assumed genetic model in association studies, especially in GWAs and other array-based studies. For this project, an achievable sample size of 475 was chosen for both control and case groups.

2.3.2.2 SNP array

Genomic DNA samples were aliquoted and transported on dry ice to the Genetics core facility (Wellcome Trust Clinical Research Facility (WTCRF), Edinburgh) within low profile 96 well PCR plates. A subset of 32 biallelic SNPs was rationally

selected for the SNP array (See **Chapter 3** for the selection of candidate SNPs). The array was performed by the WTCRF Genetics core using the QuantStudio 12K Flex Real Time PCR System, following TaqMan® OpenArray® Genotyping SOP version 2. The raw data was analyzed using TaqMan Genotyper Software v1.3 (ThermoFisher Scientific, Paisley, UK) by two experienced research technicians from the Genetics core facility (see Figure 4 for a representative allelic discrimination plot from the array) and then a genotyping report was generated. Allele and genotype frequencies of each SNP were obtained by direct counting and summarized in contingency tables. Hardy-Weinberg Equilibrium (HWE) were assessed for each SNP in both the control and case group using the Michael H. Court's (2005–2008) online calculator (Moscicki et al., 2010). For all the SNPs that passed HWE test (with a p value >0.05), Pearson's $\chi 2$ test (if all expected counts are $\geq$ 5) and Fisher's exact test (If any expected counts are <5) were performed by GraphPad Prism 5 (GraphPad Software, San Diego, CA) to assess whether the distribution of alleles differ between cases and controls under the log-additive genetic model. Raw p values obtained were further corrected for multiple comparisons through the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) using p.adjust package in R. SNPs with a FDR value < 0.15 were considered significant.

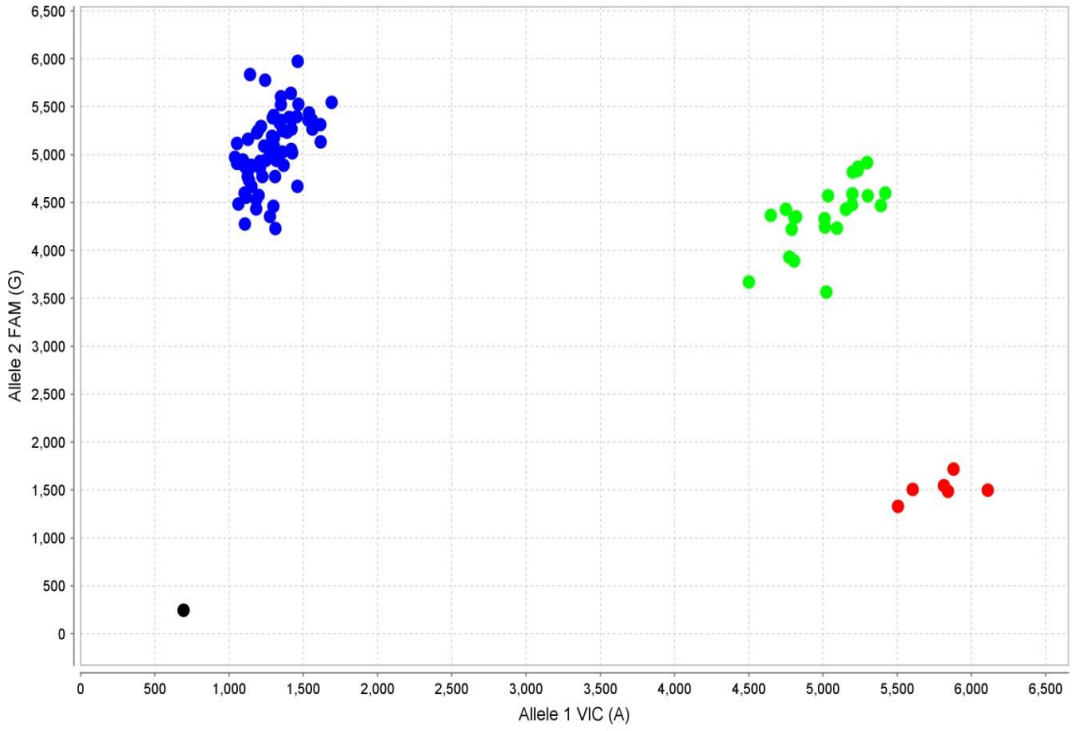


Figure 4 Representative allelic discrimination plot for a biallelic SNP

The raw data were imported into TaqMan genotyper software as allelic discrimination plots, which showed the genotype of each sample for each SNP. In this figure, each dot stands for a sample. Homozygote (blue or red) has the signal from only one probe while heterozygote (green) has signals from both probes. The black dot is the blank control on that plate. This figure showed genotyping data for rs1800629 in the first run of the SNP array.

2.3.3 Validation of the SNP array

To further validate the significant results of the SNP array, significant SNPs in the initial analysis were re-genotyped in the same set of samples using single tube TaqMan® SNP genotyping assay and LightSNiP assay with melting curve analysis. The single tube TaqMan® SNP genotyping assay was performed in the Genetics core facility following the manufacturer's instructions whilst the LightSNiP assay was performed by me in the Queen's Medical Research Institute. In the LightSNiP

assay, PCR amplification was performed in a 10µl reaction volume on the LightCycler® Nano Instrument in Strips of 8 low profile PCR tubes (See table 5 and 6 for details).

Table 5 Standard PCR reaction setup for LightSNiP assay

PCR Component	Volume
LightCycler® FastStart Reaction Mix HybProbe,	*0.5 µl*
LightCycler® FastStart Enzyme	*0.5 µl*
25mM $MgCl_2$	*0.4 µl*
PCR-grade water.	*6.1µl*
diluted gDNA (~10ng/µl)	*2 µl*
Primer/probe Mix	*0.5 µl*
Total reaction volume	*10 µl*

Table 6 Standard thermal profile for LightSNiP assay

Step	Thermal profile
1.Polymerase activation	*95°C for 10 min*
2.Denature	*95°C for 10 sec*
3.Annealing	*60°C for 15 sec*
4.Extension	*72°C for 20 sec*
5.Repeat steps 2–4	*44 times (45 in total)*
6. generate Melting curve(Extra cycle)	*30s at 95°C, 45°C to 75°C at 0.2°C*
7. Stop	*4°C forever*

A negative DNA control consisting of PCR-grade water was included in all runs in order to ensure that no sample contamination had occurred. Melting peaks were generated by plotting negative derivative of fluorescence against temperature and the melting temperature (Tm) of the products was determined. For CXCR1 rs2234671, melting peaks at 53.5 ℃ and 64 ℃ correspond to the presence of G and C allele respectively. For SULF1 rs2623047, melting peaks at 57 ℃ and 65 ℃ correspond to the presence of A and G respectively (see Figure 5 and 6 for the representative dissociation curves). The genotype of each sample was determined directly according to the melting peak profile and compared with the results obtained in the SNP array. For the purpose of quality control, the genotypes were read by at least two independent group members. Any unrecognizable results were repeated until 100% concordance was reached.

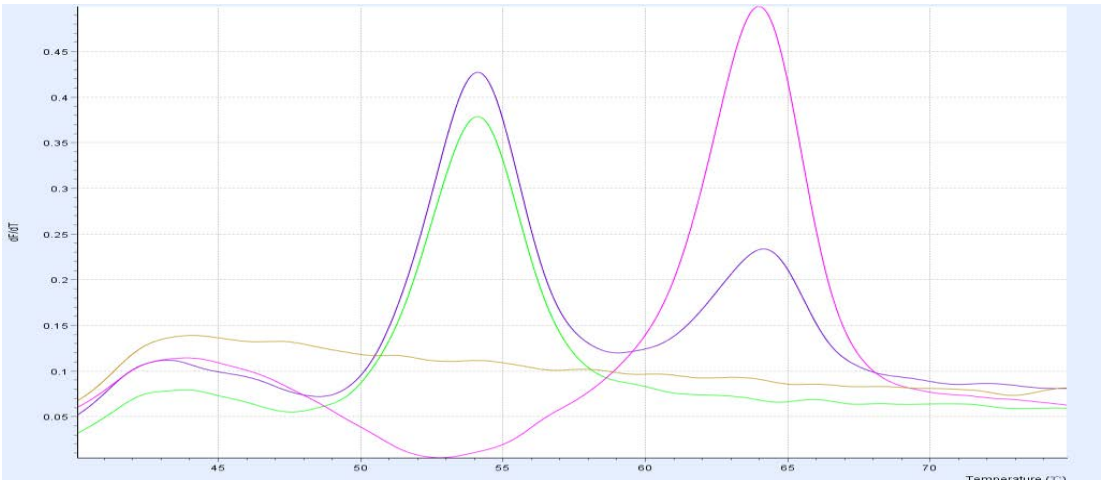


Figure 5 Representative dissociation curve of the CXCR1 rs2234671 PCR product

The figure shows the melting curves converted to melting peaks by plotting the negative derivative of fluorescence against temperature. The melting temperatures (Tm) are 53.5°C for G and 64 °C for C allele, respectively. In this figure, each curve stands for a sample. Green line shows a homozygous carrier of the G allele while pink line shows a homozygous carrier of the C allele. Purple line represents a G/C heterozygote. Negative control (PCR-grade water) was shown with the amber line.

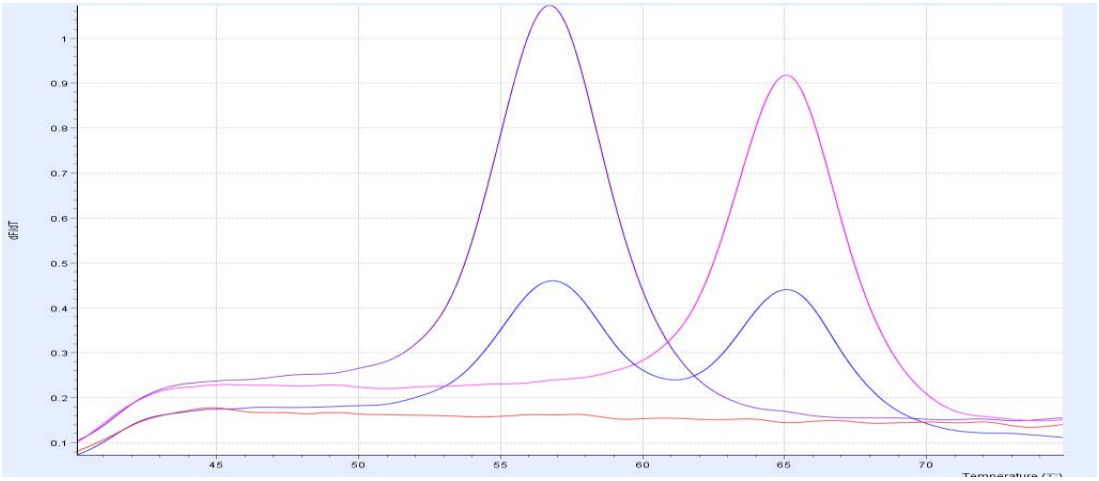


Figure 6 Representative dissociation curve of the SULF1 rs2623047 PCR product

The figure shows the melting curves converted to melting peaks by plotting the negative derivative of fluorescence against temperature. The melting temperatures (Tm) are 57 °C for A and 65 °C for G allele, respectively. In this figure, each curve stands for a sample. Pink line shows a homozygous carrier of the G allele while purple line shows a homozygous carrier of the A allele. Blue line represents an A/G heterozygote. Negative control (PCR-grade water) was shown with the red line.

In order to investigate the role of significant SNPs in a broader spectrum of HPV-related disease, genotyping results of an additional 83 HPV positive cytologically normal LBC cervical samples, 21 HPV positive cervical cancer samples, 129 HPV positive VIN samples and 23 HPV positive vulval cancer samples were combined with the validated array results. Association between each SNP and HR-HPV-related diseases was tested under five different genetic models (i.e. co-dominant, dominant, recessive, over-dominant and log-additive) using the SNPstat online tool (Solé et al., 2006). Odds ratios (OR), 95% confidence intervals (CI) were calculated for each inheritance model after age adjustment. The best–fit model for each SNP was determined according to the likelihood ratio test p value,

Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores.

2.3.4 Generation of CXCR1 mutations by Site Directed Mutagenesis

2.3.4.1 Determination of allele present in the CXCR1 cDNA clone

A cDNA clone for human CXCR1 was previously obtained from Mammalian Gene Collection (MGC) and cloned into the pDONR223 vector in Prof Jürgen Haas's lab. For verification, the vector containing CXCR1 gene was sequenced by Source BioScience, Nottingham, UK with pDONR223 forward primer 5'-TCGCGTTAACGCTAGCATG-3' and pDONR223 reverse primer 5' TAATACGACTCACTATAGGG 3'. The sequencing result was aligned with CCDS sequence data 2409.1 using EMBOSS Needle nucleotide alignment tool.

CCDS sequence data 2409.1 starts from the initiation of the ATG codon in exon 2 of the CXCR1 gene and terminates with TGA codon, containing a total of 1053 nucleotides. The major allele G of CXCR1 rs2234671 was found 827nt downstream from the start codon of the CCDS sequence data 2409.1 and therefore commonly designated as CXCR1 827G in published studies (Bellamy, 2003; Kato et al., 2000; Ragnarsdóttir et al., 2011). The minor allele C was correspondingly named as CXCR1 827C. For simplicity of notation, I also used CXCR1 827G and 827C throughout my project. This non-synonymous substitution results in an amino acid change from serine (AGC) to threonine (ACC) at the 276th amino acid residue of the CXCR1 protein.

The MGC CXCR1 cDNA clone matches the CCDS sequence data 2409.1 at nucleotide position 827. However, the two sequences have differences at three other nucleotide positions, which are 16, 92 and 1003 respectively. The MGC CXCR1 cDNA clone possesses 16T, 92G and 1003T, of which the first one is an unreported missense mutation while the other two are both minor alleles of known SNPs (rs16858811 and rs16858808 separately). Based on SNAP analysis (Johnson et al., 2008), these two SNPs are found in complete linkage disequilibrium (D'=1 and R^2=1). Genotyping results of rs16858811, rs2234671 and rs16858808 were obtained from the 1000 Genomes Project Phase 3 European database (Consortium, 2015). These three SNPs formed only three haplotypes among a total of 503 European subjects (See Table 7 for haplotype frequencies). The minor allele of rs2234671 only appeared with the major alleles of the other two SNPs. In view of the fact that the frequency of the G-G-T haplotype is quite low in the European population, I decided to change 16T, 92G and 1003T into corresponding major alleles so as to focus on the functional effects caused by rs2234671.

Table 7 CXCR1 Haplotype frequencies in 1000 Genomes Project Phase 3 European super population

Haplotype(rs16858811-rs2234671-rs16858808)	Frequency
T-G-C	*92.94% (=935/1006)*
T-C-C	*3.38% (=34/1006)*
G-G-T	*3.68% (=37/1006)*

2.3.4.2 Generation of a point mutation within the pDONR223 CXCR1 construct using Site-directed mutagenesis

2.3.4.2.1 Single site directed mutagenesis PCR

pDONR223 CXCR1 construct with C, instead of G, at nucleotide position 827 was firstly generated by site-directed mutagenesis using KOD Hot Start DNA Polymerase kit, with the forward primer 5'-GTGATCCAGGAGACCTGTGAGCGCCGC-3' and reverse primer 5'-GCGGCGCTCACAGGTCTCCTGGATCAC-3' designed by the QuikChange® Primer Design Program (Agilent Technologies, Santa Clara, CA). PCR amplification was performed in a 50µl volume reaction on GeneAmp® PCR System 9700 (see Table 8 and 9 for details).

Table 8 Standard PCR reaction setup for single site directed mutagenesis

PCR Component	Volume
10X Buffer for KOD Hot Start DNA Polymerase	*5 µl*
2mM dNTPs	*5 µl*
25mM $MgSO_4$	*4 µl*
DMSO	*1 µl*
template DNA(100ng/µl)	*1 µl*
KOD Hot Start DNA Polymerase (1U/µl)	*1 µl*
Sense (5') Primer (10 µM)	*1 µl*
Anti-Sense (3') Primer (10 µM)	*1 µl*
PCR-grade H_2O	*31 µl*
Total reaction volume	*50 µl*

Table 9 Standard thermal profile for single site directed mutagenesis.

Step	Thermal profile
1.Polymerase activation	*94°C for 5 min*
2.Denature	*94°C for 15 sec*
3.Annealing	*60°C for 30 sec*
4.Extension	*72°C for 59 sec*
5.Repeat steps 2–4	*17 times (18 in total)*
6.Stop	*72°C for 5min and 4°C forever*

2.3.4.2.2 PCR product purification

After PCR, 1µl DpnI was added into the PCR reaction and incubated for 90min at 37°C in the Eppendorf Thermomixer so as to remove parental methylated DNA and hemi-methylated DNA. The PCR product was further purified using QIAprep Spin Miniprep Kit. Briefly, the PCR product was transferred to the spin column provided by the kit and centrifuge at 15682G for 60s to discard the flow-though. The silica membrane of column was then washed with 750µl buffer PE at 15682G for another 60s. Finally, the purified PCR product was eluted in 20µl PCR grade water.

2.3.4.2.3 Diagnostic Restriction Enzyme Digestion

5µl of eluted DNA was transformed into chemically competent *Escherichia coli DH10β* as described in **Section 2.3.1.3** and then streaked onto an LB agar plate (containing 50µg/ml spectinomycin). 3 colonies were randomly selected, purified with QIAprep Spin Miniprep Kit as described in **Section 2.3.1.6.1**. According to the restriction map for pDONR223 (see Figure 7), XbaI and NheI were selected to digest

the Miniprep product at 37℃ in the thermomixer for 60min (see Table 10 for details).

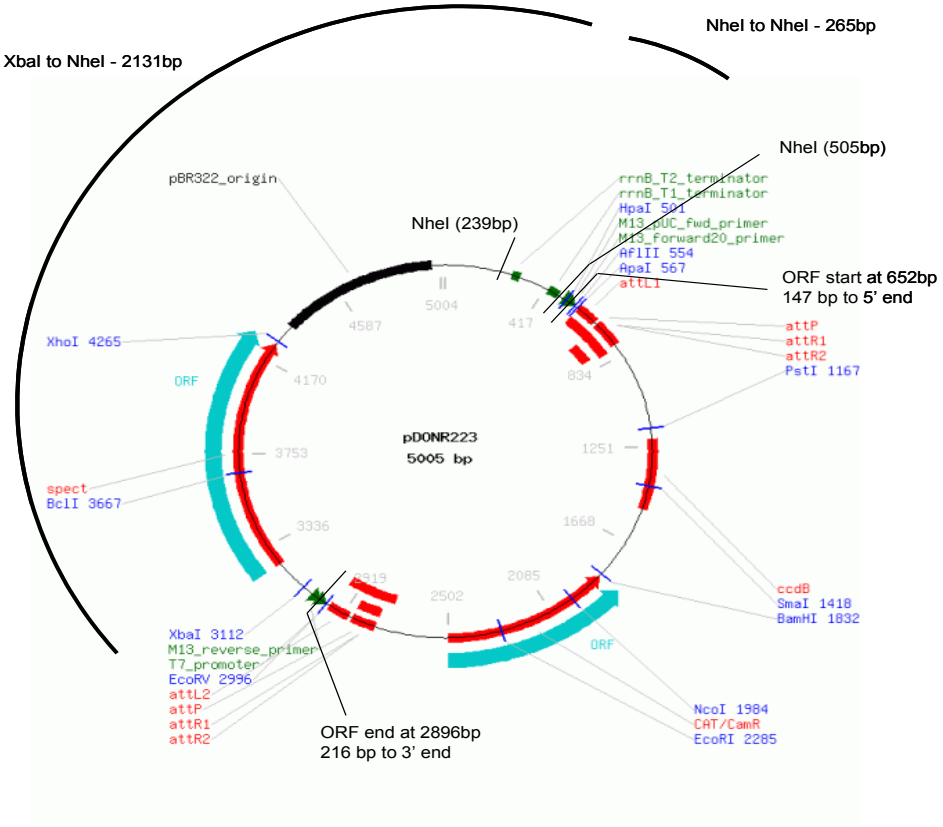


Figure 7 Restriction map of pDONR223

The diagram was provided by Dr Samantha J. Griffiths, Division of Pathway Medicine, University of Edinburgh. cDNA clone for human CXCR1 was inserted between attL1 and attL2.

Table 10 Standard fast double digestion reaction set-up for Miniprep product

Component	Volume
Enzyme 1	*0.2µl*
Enzyme 2	*0.2µl*
FastDigest Green Buffer (10X)	*2µl*
10% Bovine Serum Albumin	*2µl*
ddH_2O	*13.6µl*
Miniprep product	*2µl*

The digests were analyzed by gel electrophoresis as described in **Section 2.3.1.7** (see Figure 8 for a representative gel picture of the CXCR1 gene in pDONR223). Plasmids with the correct insert size were sequenced to check for point mutation. As expected, the C allele was discovered at nucleotide position 827 while the rest of the sequence remained unchanged.

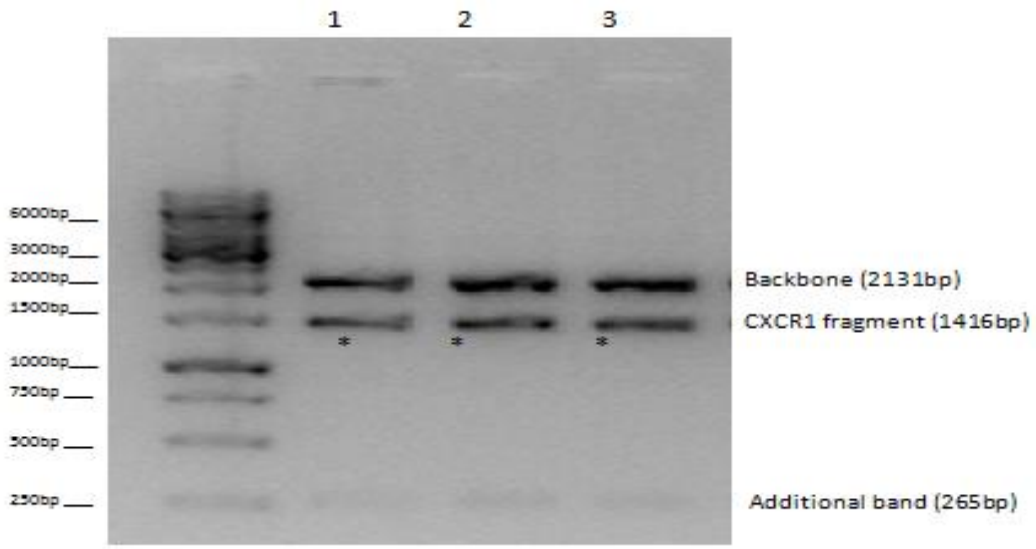


Figure 8 Restriction digest of the CXCR1 pDONR223 plasmid with XbaI and NheI

Restriction fragments were run on 1 % TAE agarose gel. SYBR® Safe DNA gel stain was used for visualization of DNA under UV light. A backbone of 2131bp, an additional band 0of 265bp (NheI to NheI) and CXCR1 (1053bp) with an extra 147bp at the 5' end and 216bp at the 3' end can be seen for all samples (lanes 1-3). CXCR1 fragment in each lane are indicated with *. DNA molecular weight marker is shown next to the lane 1, with the band sizes shown on the left.

2.3.4.2.4 Correction for the other unexpected mutations using multiple site directed mutagenesis PCR

pDONR223 CXCR1 827G or 827C construct with 16G, 92T and 1003C was generated by multiple site directed mutagenesis using KOD Hot Start DNA Polymerase kit. Three pairs of primers were designed by the QuikChange® Primer Design Program (See Table 11 for details of the primer sets) and used simultaneously in the PCR reaction. The reaction was performed in a 50µl volume on GeneAmp® PCR System 9700 (see Table 12 for the reaction set-up) with the same thermal profile as single site directed mutagenesis (see Table 9 for the detailed thermal profile).

Table 11 Primer sets used for multiple site directed mutagenesis

Site/Primer type	Primer sequence
T16G/Forward primer	*5'-atgtcaaatattacagatccacagatgtgggattt-3'*
T16G/Forward primer	*5'-aaatcccacatctgtggatctgtaatatttgacat-3'*
G92T Forward primer	*5'-gaagattacagcccctgtatgctagaaactgagaca-3'*
G92T Reverse primer	*5'-tgtctcagtttctagcatacaggggctgtaatcttc-3'*
T1003C Forward primer	*5'-gagttcttggcacgtcatcgtgttacctcctaca-3'*
T1003C Reverse primer	*5'-tgtaggaggtaacacgatgacgtgccaagaactc-3'*

Table 12 Standard PCR reaction setup for multiple site directed mutagenesis

PCR Component	Volume
10X Buffer for KOD Hot Start DNA Polymerase	*5 µl*
2mM dNTPs	*5 µl*
25mM $MgSO_4$	*4 µl*
DMSO	*1 µl*
template DNA(100ng/µl)	*1 µl*
KOD Hot Start DNA Polymerase (1U/µl)	*1 µl*
Sense (5') Primer (10 µM)	*1 µl/ set×3*
Anti-Sense (3') Primer (10 µM)	*1 µl/ set×3*
PCR-grade H_2O	*27 µl*
Total reaction volume	*50 µl*

PCR products for pDONR223 CXCR1 827G or 827C were purified and transformed into *Escherichia coli DH10β* as described in Section 2.3.4.2.2 and 2.3.1.3. 12 colonies were selected from each plate and cultured in 96-well deep well plate with 2ml/well LB medium containing 50µg/ml spectinomycin. Plasmid DNA was then extracted using alkaline lysis method as described in Section 2.3.1.6.2. According to the restriction map for pDONR223 (see Figure 7), XbaI and NheI were selected to digest the alkaline lysis product at 37℃ in the thermomixer for 60min (see Table 13 for details). For each construct, 6 out of 12 digests were randomly picked and analyzed by gel electrophoresis as described in Section 2.3.1.7. All digests were found with correct insert size and therefore were all sent out for sequencing. Plasmids with major allele at all three nucleotide sites (i.e. 16, 92 and 1003) were selected for the following experiments.

Table 13 Standard fast double digestion reaction set-up for alkaline lysis product

Component	Volume
Enzyme 1	*0.2µl*
Enzyme 2	*0.2µl*
FastDigest Green Buffer (10X)	*2µl*
10% Bovine Serum Albumin	*2µl*
ddH_2O	*4.6µl*
RNase A(10mg/ml)	*1µl*
Alkaline lysis product	*10µl*

2.3.5 Establishment of cell based models for CXCR1 rs2234671

2.3.5.1 Transient Transfection model

2.3.5.1.1 Transfer CXCR1 gene from pDONR223 vector to pCR3 destination vector

After introducing the point mutation into the CXCR1 gene, both CXCR1-827G and CXCR1-827C genes were cloned from pDONR223 Gateway® Donor vector into pCR3 destination vector using LR recombination reaction according to the manufacturer's instructions (Gateway cloning system, Invitrogen, UK). The pCR3 destination vector containing the Gateway® cassette at an EcoRV site was previously generated in Prof Jurgen Haas's lab. In brief, 1.5µl pDONR223 with CXCR1 insert, 1µl pCR3 destination vector (300ng/µl), 1µl of LR clonase® II enzyme and 1µl double distilled H2O (ddH_2O) were mixed together and incubated at room temperature (RT) overnight. 1µl of this reaction mixture was then used to transform *Escherichia coli DH10β* as described in **Section 2.3.1.3**. Bacteria were

subsequently streaked onto LB agar plates (containing 100µg/ml Ampicillin). 10 Colonies were randomly selected from each plate, purified as described in Section 2.3.1.6.1 and restriction-digested with EcoRV. The digests were analyzed by gel electrophoresis as described in Section 2.3.1.7 (see Figure 9 for a representative gel picture of the CXCR1 gene in pCR3 vector).

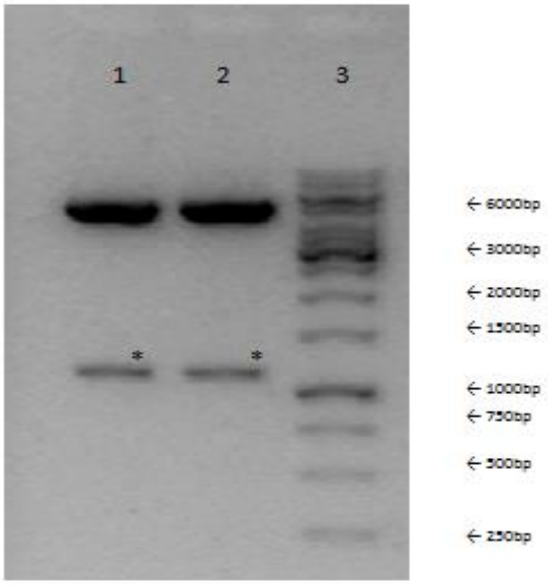


Figure 9 Restriction digest of the CXCR1 pCR3 plasmid with EcoRV

Digests were run on 1 % agarose gel. SYBR® Safe DNA gel stain was used for visualization of DNA under the UV light. A backbone of 5.1kb and CXCR1 fragment (1053bp) with an extra 124bp at both 5' end and 3' end can be seen for both 827G and 827C plasmids (lanes 1 and 2, respectively). CXCR1 fragment in each lane are indicated with *. DNA molecular weight marker is shown in lane 3, with the band sizes shown on the right.

2.3.5.1.2 Transient transfection of HeLa cells and HEK293 cells

HeLa cells or HEK293 cells were seeded in antibiotic free growth medium 24 hours before transfection. Please see Table 14 for optimal seeding density of HeLa cells and HEK293 cells under different conditions.

Table 14 Optimal seeding density of HeLa and HEK293 cells

Culture plates	HeLa cells	HEK293 cells
6 well plate	*5×10^5/well*	*6.25×10^5/well*
24 well plate	*1×10^5/well*	*1.25×10^5/well*
96 well plate	*2×10^4/well*	*2×10^4/well*
8 well chamber slide	*6×10^4/well*	*7×10^4/well*

Plasmids to be transfected were diluted in Opti-MEM I Reduced Serum Medium and mixed with diluted lipofectamine 2000 according to manufacturer's instructions. Reaction volume was adjusted according to the size of the culture (see Table 15 for details).

Table 15 Lipofectamine® 2000 DNA transfection reagent protocol for HeLa and HEK293 cells

Culture plates	*HeLa cells (per well)*		*HEK293 cells (per well)*	
	Plasmid	*Lipofectamine® 2000*	*Plasmid*	*Lipofectamine® 2000*
6 well plate	*2500ng*	*12.5ul*	*2500ng*	*7.5ul*
24 well plate	*500ng*	*2.5ul*	*500ng*	*1.5ul*
96 well plate	*100ng*	*0.5ul*	*100ng*	*0.3ul*
8 well chamber slide	*300ng*	*1ul*	*300ng*	*0.9ul*

The plasmid-lipid mixture was incubated at RT for 20 minutes and then added directly onto cells. For HeLa cells, medium was changed for antibiotic free growth medium 6h after transfection to avoid cell toxicity. However, medium did not need to be changed for HEK293 cells within 24h.

2.3.5.1.3 CXCR1 expression in HEK293 cells by fluorescence microscopy

HEK293 cells were plated in 96-well plate and transfected with one of the CXCR1 expressing plasmid (pCR3-CXCR1 827G or pCR3-CXCR1 827C) or the vector alone (pCR3) using Lipofectamine 2000 as described in **Section 2.3.5.1.2**. The next day, cells were fixed with 10% neutral buffered Formalin solution at RT for 10min and blocked with 1% Bovine Serum Albumin (BSA) in PBS at RT for 1h. Finally, cells were stained with APC Anti-human CD181[7] antibody (1 in 200 diluted in 1% BSA blocking buffer) in dark at RT for 30min. After rinsing with PBS, cells were examined at 10X magnification under the EVOS™ digital inverted microscope. Figure 10 shows representative images of the CXCR1 expression in transfected HEK293 cells.

7 CXCR1 has also been designated as CD181

A) **HEK293 cells transfected with pCR3 vector**

B) **HEK293 cells transfected with pCR3-CXCR1 827G**

C) **HEK293 cells transfected with pCR3-CXCR1 827C**

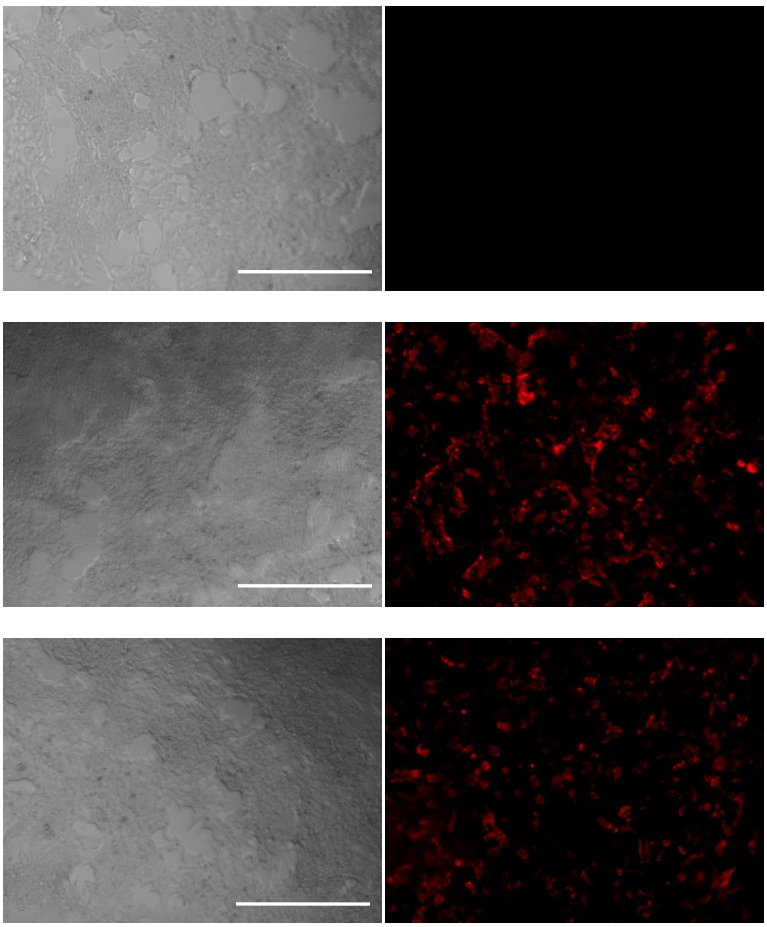


Figure 10 CXCR1 is expressed in pCR3-CXCR1 827G and 827C transfected HEK293 cells

Picture A: HEK293 cells transfected with pCR3 (Left: view visualised under the transmitted light cube, Right: exact same field visualised under the Cy5 cube); Picture B: HEK293 cells transfected with pCR3-CXCR1 827G (Left: view visualised under the transmitted light cube, Right: the exact same field visualised under the Cy5 cube); Picture C: HEK293 cells transfected with pCR3-CXCR1 827C (Left: view visualised under the transmitted light cube, Right: exact same field visualised under the Cy5 cube); All pictures were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) and the scale bar=400µm.

2.3.5.1.4 CXCR1 expression in HeLa cells by confocal microscopy

HeLa cells (1.2×10^5/well) were plated in 8-well chamber slide 24h before transfection. The next day, cells were transfected with CXCR1 expressing plasmid (pCR3-CXCR1 827G or pCR3-CXCR1 827C), green fluorescent protein expressing plasmid (pCR3-GFP) or vector alone (pCR3) using Lipofectamine 2000 according to the manufacturer's instructions (with a final concentration of 300ng DNA and 1μl Lipofectamine 2000 reagent /well). The medium was changed 6h after transfection and the cultures were maintained for an additional 24h. Then, cells were fixed with 10% neutral buffered formalin solution at RT for 10min, permeabilized with 0.5% Triton ™ X-100 at RT for 5min, blocked with 1% BSA in PBS at RT for 1h. Finally, cells were stained with APC-labelled anti-human CXCR1 antibody (1 in 200 diluted in 1% BSA blocking buffer) in dark at RT for 30min. After rinsing thrice with PBS and ddH_2O, the chamber was removed and the slide was mounted with ProLong® Gold Antifade Mountant with DAPI and observed under a Leica TCS SP5 confocal microscope. Figure 11 shows representative pictures of CXCR1 expression on transfected HeLa cells.

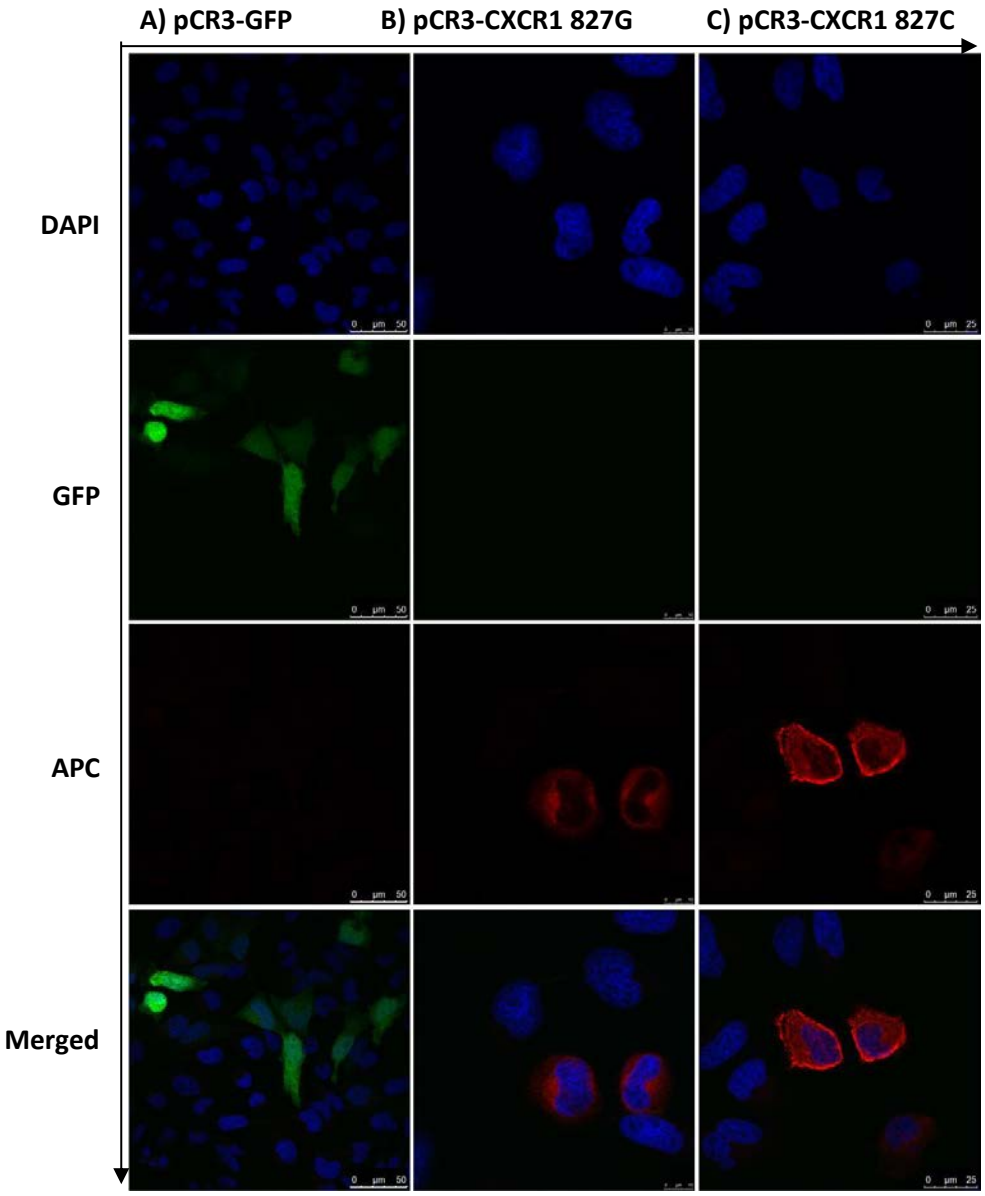


Figure 11 Expression of CXCR1 in pCR3-CXCR1 827C or 827G transfected HeLa cells

HeLa cells were transiently transfected with pCR3-GFP, pCR3-CXCR1 827G or pCR3-CXCR1 827C in an 8-well chamber slide. 30h after transfection, cells were fixed, permeabilized and stained with APC-conjugated mouse anti-human CD181 (CXCR1) antibody. The slide was mounted using ProLong® Gold Antifade Mountant with DAPI. All images were obtained by confocal microscopy. In each condition, representative images of the same field of view visualised under different channels (i.e. DAPI, GFP, APC and Merged) were shown. (A) Cells transfected with pCR3-GFP were stained negative for CXCR1. (B) Cells transfected with pCR3-CXCR1 827G were stained positive for CXCR1. C) Cells transfected with pCR3-CXCR1 827C were stained positive for CXCR1.The scale markers represent 50µm in A, 10µm in B and 25µm in C.

2.3.5.2 Stably transduced cell lines

2.3.5.2.1 Transfer of CXCR1 gene from pDONR223 vector to pLenti6/V5-DEST® Gateway® vector

CXCR1 827G and CXCR1 827C genes were cloned from pDONR223 Gateway® Donor vector into pLenti6/V5-DEST™ Gateway® Vector using LR recombination reaction according to the manufacturer's instructions. pLenti6/V5-DEST™ destination vector contains Gateway® cassette between attR1 and attR2 sites. In brief, 1µl pDONR223 with CXCR1 insert, 1µl pLenti6/V5-DEST™ Gateway® Vector (300ng/µl), 2µl of LR clonase® II enzyme and 1µl double distilled H2O (ddH_2O) were mixed together and incubated at RT (RT) overnight. 2µl of this reaction mixture was then used to transform electrocompetent Escherichia coli *Stbl3™* as described in **Section 2.3.1.4.** Bacteria were subsequently streaked onto LB agar plates (containing 100µg/ml Ampicillin). 10 Colonies were randomly selected from each plate, purified as described in **Section 2.3.1.6.1** and restriction-digested with EcoRV. The digests were analyzed by gel electrophoresis as described in **Section 2.3.1.7** Figure 12 shows a representative gel picture of CXCR1 gene in pLenti6/V5-DEST™ vector.

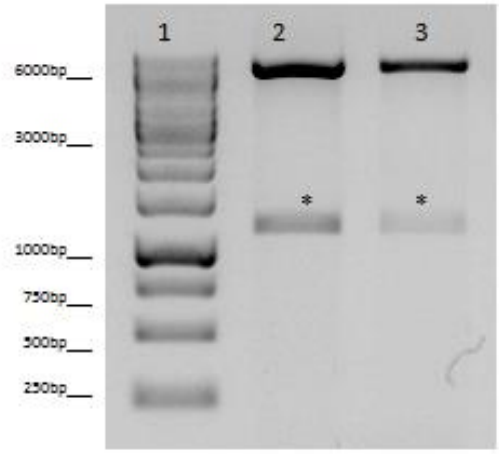


Figure 12 Restriction digest of CXCR1 pLenti6/V5-DEST™ plasmid with EcoRV

Digests were run on 1 % agarose gel. SYBR® Safe DNA gel stain was used for visualization of DNA under the UV light. A backbone of 6.9kb and CXCR1 fragment (1053bp) with an extra 129bp at the 5' end and 128bp at the 3' end can be seen for both 827G and 827C plasmids (lanes 2 and 3, respectively). CXCR1 fragment in each lane is indicated with *. DNA molecular weight marker is shown in lane 1, with the band sizes shown on the left.

2.3.5.2.2 Generation of lentivirus particles containing CXCR1 827G or827C

1.5×10^6 293T cells were seeded in antibiotic-free growth medium in 10cm dishes and cultured overnight at 37°C, 5% v/v CO2, 100% humidity as described in **Section 2.3.1.8**. The next day, each dish of cells was transfected with a mixture of the pLP1, pLP2, pLP/VSVG vectors, pLenti-based expression vector containing gene of interest and Lipofectamine 2000 (each dish finally contained 1µg of each plasmid and 12.5µl of Lipofectamine 2000). 24h after transfection, medium was changed for 10ml fresh antibiotic-free growth medium. Cells were kept in culture for another three days with the morphology being closely monitored. Once syncytia were formed by the majority of cells (Figure 13), supernatant was collected by centrifuging at 1600G for 15 minutes, filtered through a 0.45µm filter and aliquoted in 1ml volumes.

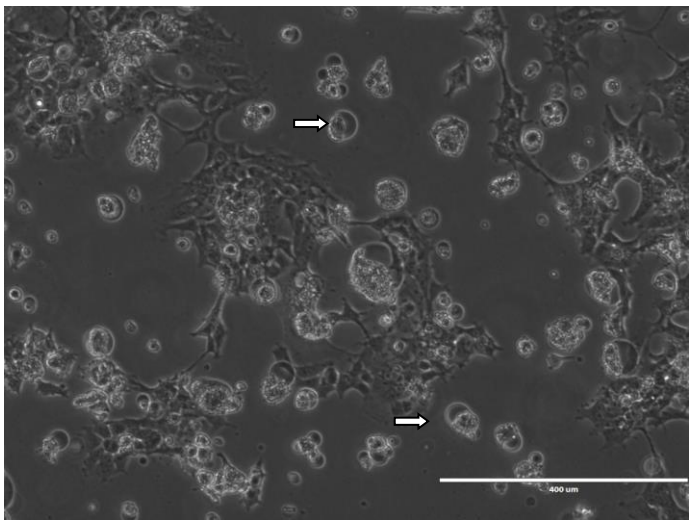


Figure 13 Syncytium formation in 293T cells producing recombinant lentiviruses

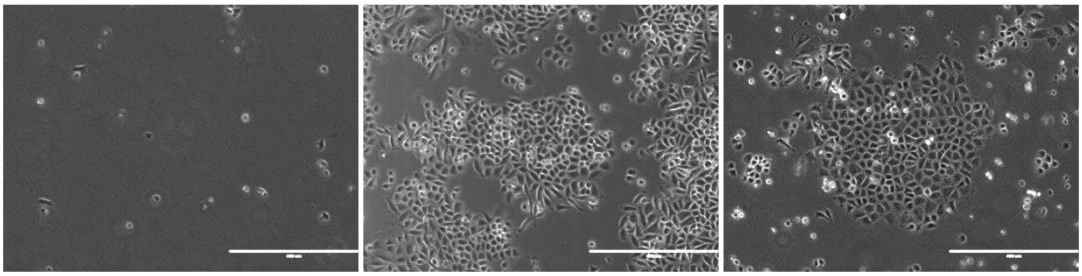
Representative picture of lentivirus producing 293T cells 3 days after transfection with ViraPower™ lentiviral products. Syncytia formed by lentivirus producing cells were highlighted with white arrows. The picture was taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) and the scale bar=400µm.

2.3.5.2.3 Determination of the minimum concentration of blasticidin used for selection

CaSki and SiHa cells were chosen for the transduction. Both cell lines were seeded in 24-well plate at a concentration of 1×10^5 cells/well. The next day, the normal growth medium was changed into medium containing a range of concentrations of blasticidin (0, 1, 2, 4, 6, 8, 10, 12µg/ml). Cells were cultured in blasticidin containing medium as described in **Section 2.3.1.8** with the viability being closely monitored. Medium was changed every two days to maintain the blasticidin concentration until the minimum concentration of blasticidin that killed all the cells after 7 days was determined for each cell line. The optimal blasticidin selection concentration was 4µg/ml and 2µg/ml for CaSki cells and SiHa cells respectively.

2.3.5.2.4 Generation of stably transduced cell lines

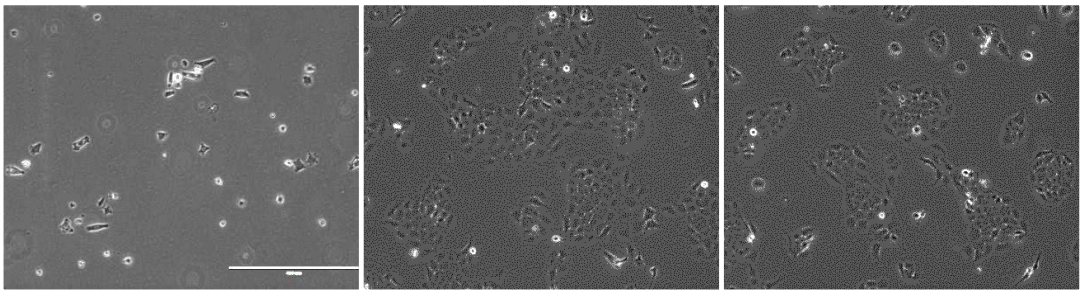
Cell lines were seeded in 12-well plates at a concentration of 2.5×10^5 cells/well. The next day, cells were transduced with 1ml supernatant containing lentivirus. Mock transduced cells were included for each cell line as controls for viability assessment. Medium was changed into normal growth medium on Day 3 and the selection was started on Day 4, with 4µg/ml blasticidin for CaSki cells and 2µg/ml blasticidin for SiHa cells. After reaching confluence, cells were split 1:4 into 6-well plates. Blasticidin containing medium was changed every two days to maintain stable selection concentration for each cell line. Once all the control cells were killed by blasticidin treatment, stably transduced cells were collected and expanded in 10cm dishes or T75 culture flasks (Figure 14 and 15 show representative pictures of cells after selection).



A) Parental SiHa **B) CXCR1 827G SiHa** **C) CXCR1 827C SiHa**

Figure 14 Transduction of SiHa cells

Representative pictures of transduced SiHa cells and mock-transduced cells under 2µg/ml blasticidin selection. Picture A: Parental SiHa cells (Mock transduced) were nearly all dead under blasticidin selection; Picture B: SiHa cells transduced with the CXCR1 827G gene survived after blasticidin selection; Picture C: SiHa cells transduced with the CXCR1 827C gene survived after blasticidin selection. All pictures were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) and the scale bar=400µm.



A) Parental CaSki **B) CXCR1 827G CaSki** **C) CXCR1 827C CaSki**

Figure 15 Transduction of CaSki cells

Representative pictures of transduced CaSki cells and mock-transduced cells under 4µg/ml blasticidin selection. Picture A: Parental CaSki cells (Mock transduced) were nearly all dead under blasticidin selection; Picture B: CaSki cells transduced with the CXCR1 827G gene survived after blasticidin selection; Picture C: CaSki cells transduced with the CXCR1 827C gene survived after blasticidin selection. All pictures were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) and the scale bar=400µm.

2.3.5.2.5 Checking the genotypes of the parental cell lines, transduced cell lines and other cervical epithelial lines using Melting curve analysis

Genomic DNA was extracted from each cell line using MasterPure™ Complete DNA and RNA Purification Kit following the manufacturer's instructions and then genotyped for rs2234671 using LightSNip assay described in **Section 2.3.3**. Figure 16 and 17 show representative genotyping results.

A) Parental SiHa B) CXCR1 827G SiHa C) CXCR1 827C SiHa

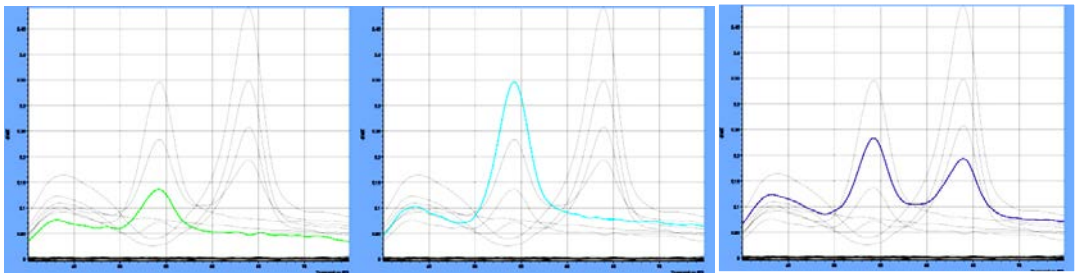


Figure 16 Genotypes of Parental SiHa cells and transduced SiHa cells

Picture A: Parental SiHa cells were G homozygotes for CXCR1 rs2234671; Picture B: SiHa cells transduced with the CXCR1 827G gene were still G homozygotes for CXCR1 rs2234671; Picture C: SiHa cells transduced with the CXCR1 827C gene became G/C heterozygotes for CXCR1 rs2234671. All pictures were generated using LightCycler ® Nano software V1.1.0.

A) Parental CaSki B) CXCR1 827G CaSki C) CXCR1 827C CaSki

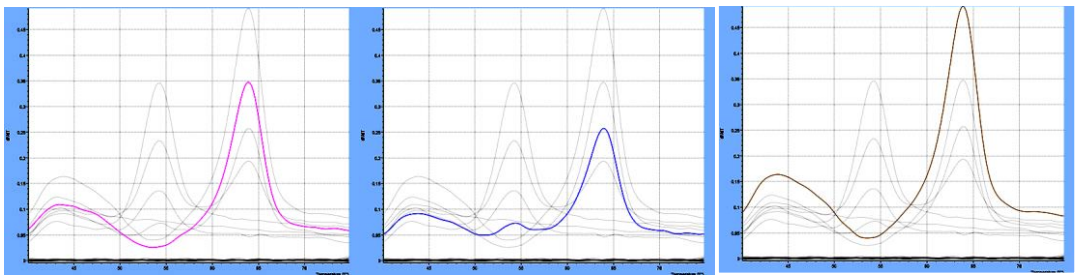


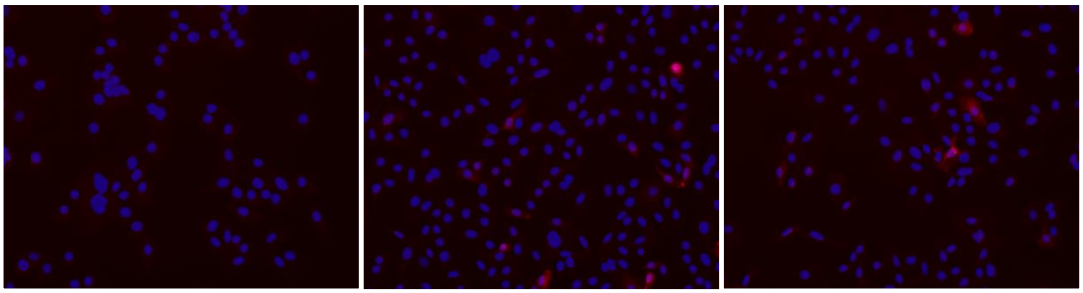
Figure 17 Genotypes of Parental CaSki cells and transduced CaSki cells

Picture A: Parental CaSki cells were C homozygotes for CXCR1 rs2234671; Picture B: CaSki cells transduced with the CXCR1 827G gene became G/C heterozygotes for CXCR1 rs2234671; Picture C: CaSki cells transduced with the CXCR1 827C gene were still C homozygotes for CXCR1 rs2234671. All pictures were generated using LightCycler ® Nano software V1.1.0.

2.3.5.2.6 Assessment of CXCR1 expression using fluorescence microscopy

Parental cells and transduced cells (1.2×10^5/well) were plated in 8-well chamber slide and cultured overnight. The next day, cells were fixed with 10% neutral buffered formalin solution at RT for 10min, permeabilized with 0.5% Triton ™ X-100 at RT for 5min, blocked with 1% BSA in PBS at RT for 1h. Finally, cells

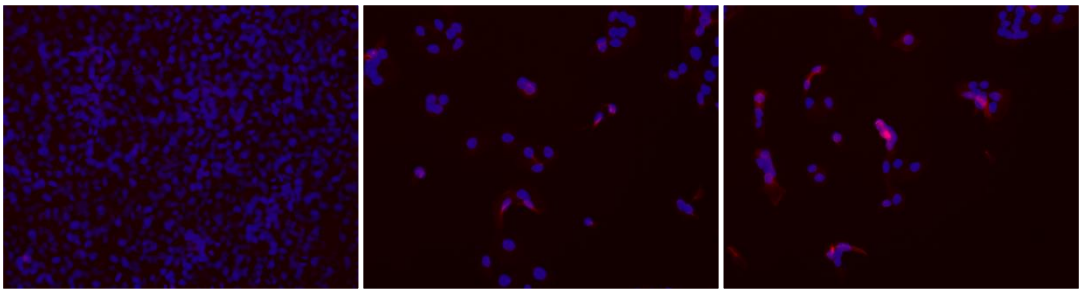
were stained with APC-labelled anti-human CXCR1 antibody (1 in 200 diluted in 1% BSA blocking buffer) in dark at RT for 30min. After rinsing thrice with PBS and ddH_2O, the chamber was removed and the slide was mounted with ProLong® Gold Antifade Mountant with DAPI and observed 40X under the EVOS™ digital inverted microscope. Figure 18 and 19 show representative pictures of CXCR1 expression.



A) Parental SiHa **B) CXCR1 827G SiHa** **C) CXCR1 827C SiHa**

Figure 18 Expression of CXCR1 in parental and transduced SiHa cells

Parental SiHa cells and SiHa cells transduced with CXCR1 827G or CXCR1 827C were seeded in an 8 well chamber slide and cultured overnight. The next day, cells were fixed, permeabilized and stained with APC conjugated mouse anti-human CD181 (CXCR1) antibody (1:200). The slide was mounted using ProLong® Gold Antifade Mountant with DAPI. All images were obtained 40X under the EVOS™ digital inverted microscope. For each cell line, representative merged image taken under both DAPI and APC channels was shown. Picture A: parental SiHa cells were not positively stained with APC anti-CXCR1; Picture B: positively APC anti-CXCR1 stained cells were detected within SiHa cells transduced with the CXCR1 827G gene; Picture C: positively APC anti-CXCR1 stained cells were detected within SiHa cells transduced with the CXCR1 827C gene.



A) Parental CaSki **B) CXCR1 827G CaSki** **C) CXCR1 827C CaSki**

Figure 19 Expression of CXCR1 in Parental and transduced CaSki cells

Parental CaSki cells and CaSki cells transduced with CXCR1 827G or CXCR1 827C were seeded in an 8 well chamber slide and cultured overnight. The next day, cells were fixed, permeabilized and stained with APC conjugated mouse anti-human CD181 (CXCR1) image antibody (1:200). The slide was mounted using ProLong® Gold Antifade Mountant with DAPI. All images were obtained 40X under the EVOS™ digital inverted microscope. For each cell line, representative merged taken under both DAPI and APC channels was shown. Picture A: parental CaSki cells were not positively stained with APC anti-CXCR1; Picture B: positively APC anti-CXCR1 stained cells were detected within CaSki cells transduced with the CXCR1 827G gene; Picture C: positively APC anti-CXCR1 stained cells were detected within CaSki cells transduced with the CXCR1 827C gene.

2.3.5.2.7 Check CXCR1 expression using flow cytometry

Cells were washed with PBS and detached from the T75 flask by either 3ml Trypsin-EDTA solution or 5ml EDTA solution. Following cell detachment, cells were pelleted by centrifuging at 190G, 4°C for 5min and further washed twice with PBS. During the wash steps, cell clumps were broken up by gently pipetting up and down. Cells were resuspended at ~1 × 10^6 cells/100µl FACS wash buffer (PBS containing 2% FBS) and transferred to a 5ml Corning™ Falcon™ Round-Bottom Polystyrene Tubes. 5µl Human TruStain FcX™ were mixed with the cells and incubated at RT for 10 min. APC Anti-human CXCR1 antibody (5ul/million cells) was then added to the cells and incubated in the dark for 30min (2–8º C). Excess antibody was removed by washing the cells with 2mL FACS wash

buffer. The final pellet was resuspended in 400µl FACS wash buffer before flow cytometric analysis with a 5 laser BD Biosciences LSR Fortessa analyser. FacsDiva software v6.1 (BD Biosciences, Oxford, UK) was used for digital data acquisition and post-acquisition analysis was performed using FlowJo software v7.6.5 (Tree Star Inc., Ashland, Oregon, USA).

During flow cytometry, debris, dead cells and doublets were excluded step by step. Cells were identified according to their Forward light scatter, FSC (proportional to cell size) and Side light scatter, SSC (proportional to cell granularity or internal complexity) and investigated for their expression of CXCR1 (Figure 20 and 21 show representative flow cytometry plots for SiHa cells and CaSki cells respectively). The expression of CXCR1 by transduced cells was determined on a univariate histogram and a cut-off fluorescence intensity value of 5×10^3 was applied to determine CXCR1 transgene expression in transduced cells. Since trypsinization did not affect CXCR1 expression on transduced cells (see Figure 22), Trypsin-EDTA solution was used to detach cells in all functional assays to improve recovery of sufficient numbers of viable cells.

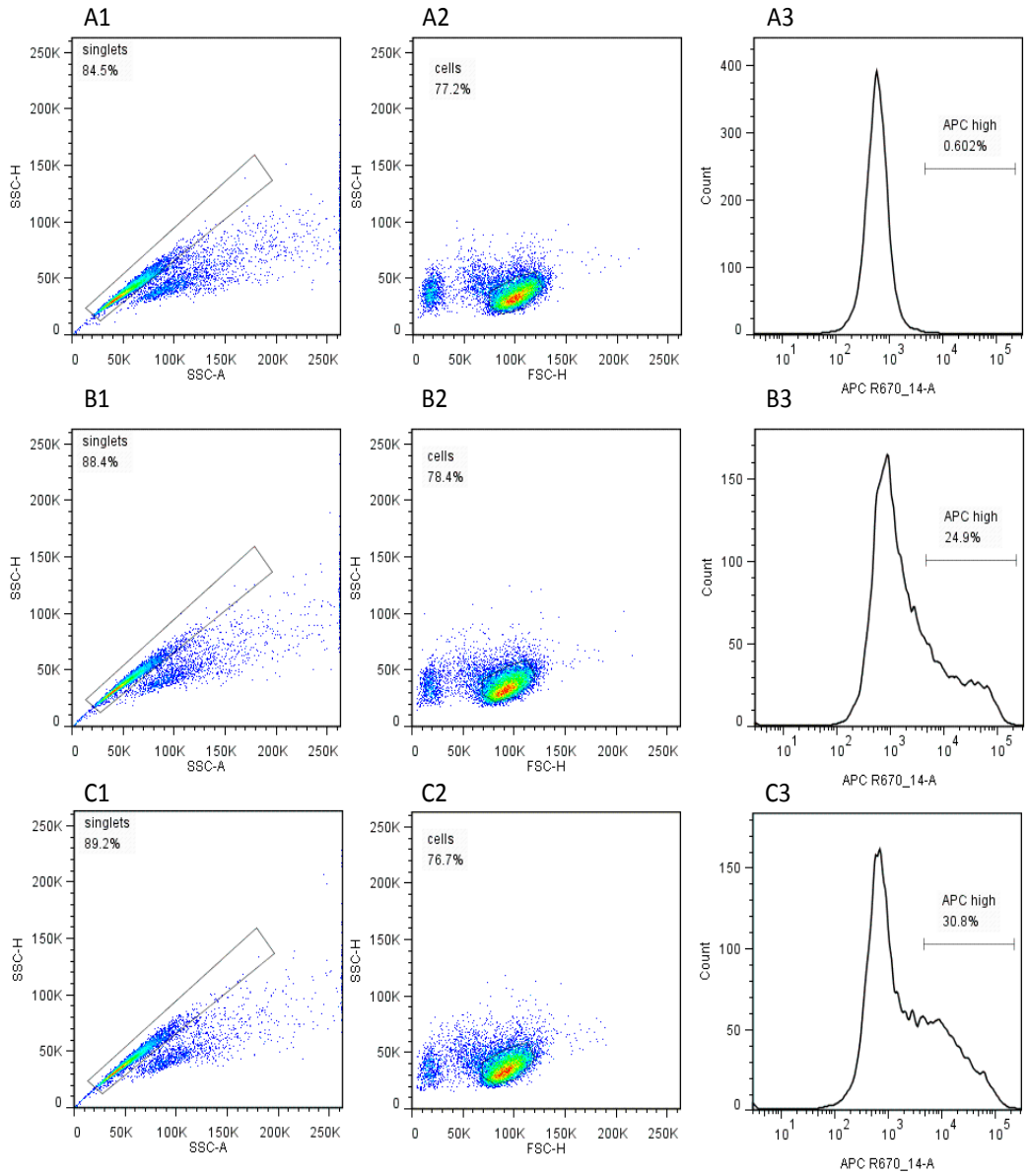


Figure 20 Representative flow cytometry plots for parental and transduced SiHa cells

A1-A3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing parental SiHa cells and histogram of CXCR1 expression by parental SiHa cells, respectively; B1-B3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827G transduced SiHa cells and histogram of CXCR1 expression by CXCR1 827G transduced SiHa cells, respectively; C1-C3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827C transduced SiHa cells and histogram of CXCR1 expression by CXCR1 827C transduced SiHa cells, respectively. By comparing with parental cells, a ranged gate is delineated in histograms to highlight transgene expression in transduced SiHa cells.

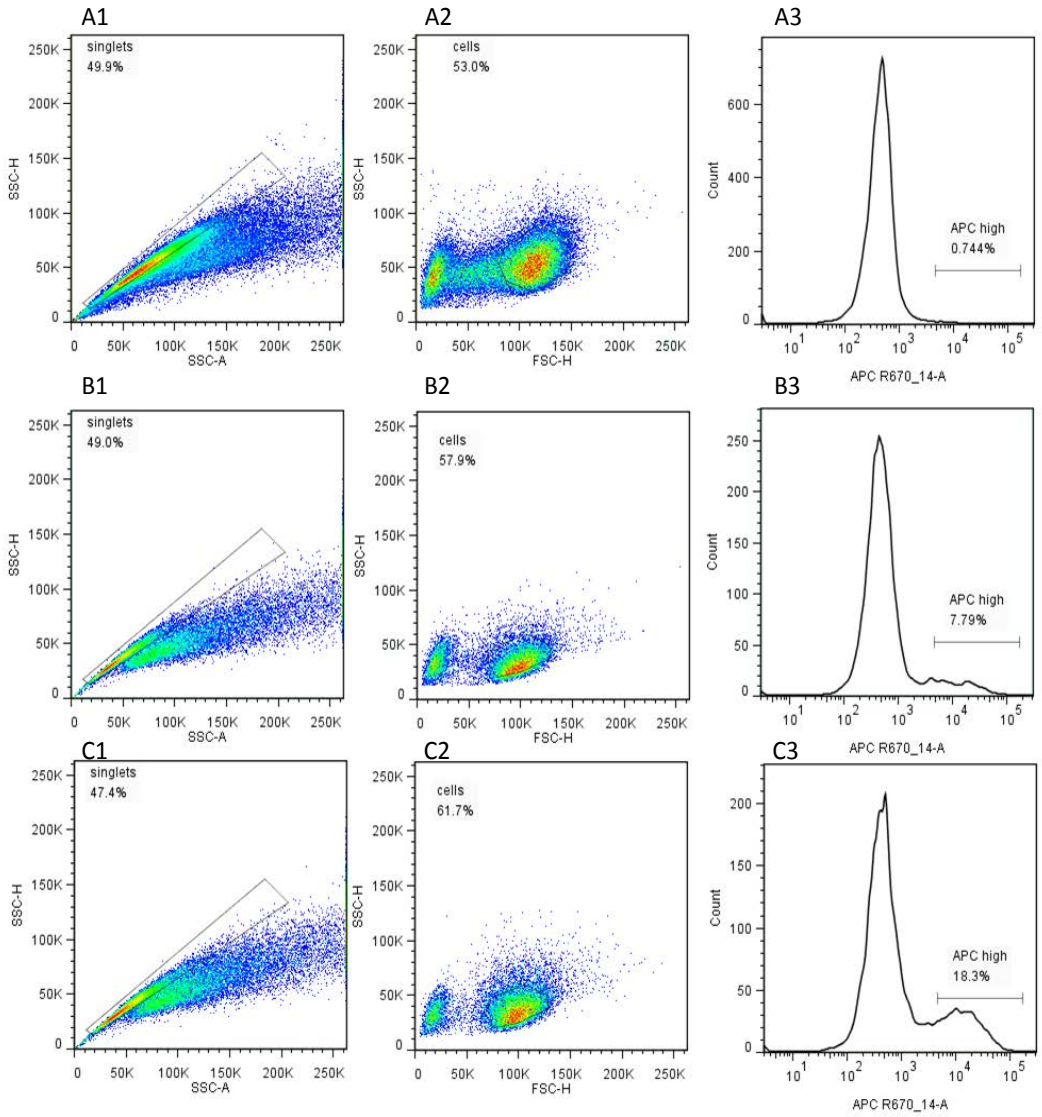


Figure 21 Representative flow cytometry plots for parental and transduced CaSki cells

A1-A3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing parental CaSki cells and histogram of CXCR1 expression by parental CaSki cells, respectively; B1-B3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827G transduced CaSki cells and histogram of CXCR1 expression by CXCR1 827G transduced CaSki cells, respectively; C1-C3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827C transduced CaSki cells and histogram of CXCR1 expression by CXCR1 827C transduced CaSki cells, respectively. By comparing with parental cells, a ranged gate is delineated in histograms to highlight transgene expression in transduced CaSki cells.

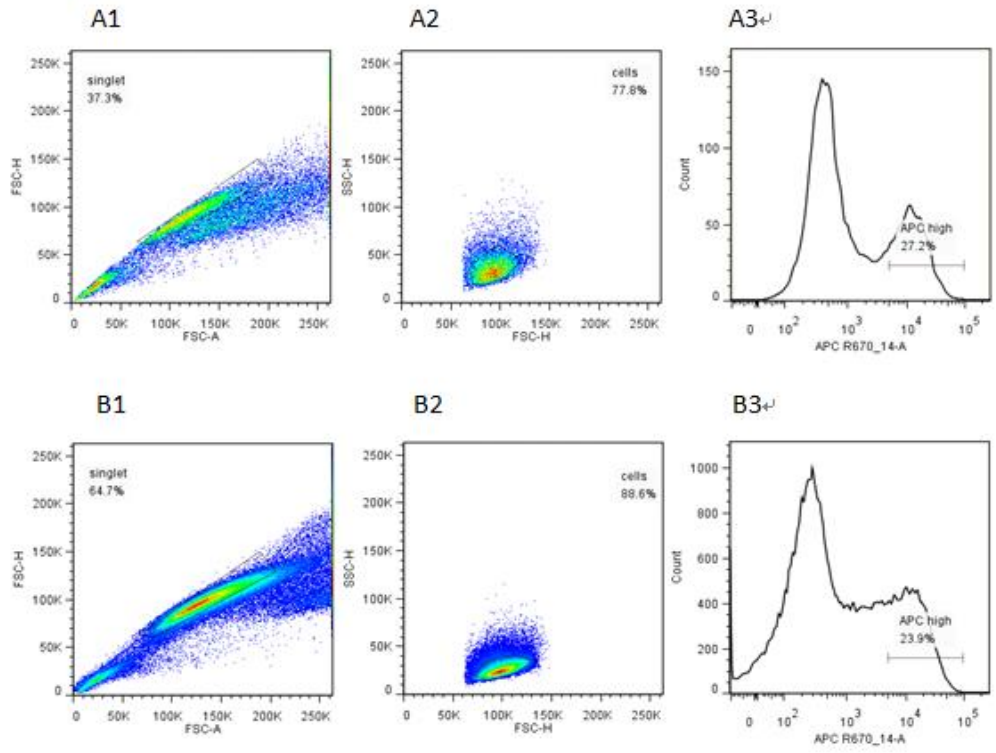


Figure 22 Trypsinization does not affect CXCR1 expression on transduced cells

A1-A3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827C transduced CaSki cells (detached with EDTA solution) and histogram of CXCR1 expression by CXCR1 827C transduced CaSki cells, respectively; B1-B3：pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827C transduced CaSki cells (detached with Trypsin-EDTA solution) and histogram of CXCR1 expression by CXCR1 827C transduced CaSki cells, respectively; FSC parameter and SSC parameters of the cells were slightly different between the two methods but CXCR1 expression was not affected.

2.3.5.2.8 Collect cells expressing high levels of CXCR1 by cell sorting

As shown in Section 2.3.5.2.7, each transduced cell line is essentially a polyclonal pool of cells stably expressing different levels of CXCR1. Since variation in gene expression could hinder our ability to investigate the functional effects caused by CXCR1 rs2234671, transduced cells expressing the highest levels of CXCR1 were sorted so as to get a relatively uniform population. Briefly, transduced cells were cultured in T162 flask as described in Section 2.3.1.8. Once reaching confluence, cells were detached from the flask using Trypsin-EDTA (0.5%) and pipetted up and down to get rid of remaining clumps. Next, cells were counted, pelleted, washed with PBS and resuspended in 300µl FACS wash buffer (PBS with 2% FBS). Then, cells were stained with 15µl APC Anti-human CXCR1 antibody in the dark for 30min. Excess antibody was removed by washing the cells with 10ml FACS wash buffer at 190G, 4℃ for 5min. Cells were further filtered onto 40µm Falcon™ Cell Strainers and resuspended in 1ml Pre-sort buffer (DMEM medium containing 1% FBS, 1% L-glutamine, 100µg/ml streptomycin and 100U/ml penicillin) at 10 million/ml. Cells were sorted on FACSAria II with a 100µm nozzle with the help of Dr Mari Pattison at the Flow Cytometry and Cell Sorting Facility in the Queen's Medical Research Institute. Sorted cells were collected in 3ml collection medium (DMEM medium containing 50% FBS, 1% L-glutamine, 100µg/ml streptomycin and 100U/ml penicillin) within Falcon™ 15mL Conical Centrifuge Tubes, spun down at 190G, 4℃ for 5min and immediately cultured in normal growth medium. Genotypes of the sorted cells were checked using LightSNiP assay as described in Section 2.3.5.2.5. Expression of CXCR1 in sorted cells was assessed using flow cytometry as described in Section 2.3.5.2.7. Mean fluorescence intensity (MFI) of CXCR1

expression was comparable between CXCR1 827G and CXCR1 C transduced cells after sorting (Figure 23 and 24 show representative flow cytometry plots for sorted SiHa cells and CaSki cells respectively). The MFI of CXCR1 expression on stably transduced cells showed little variance over time, regardless of passage number (data not shown). I also checked CXCR2[8] expression on these transduced cells. However, staining with PE anti-human CD182 (CXCR2) antibody did not produce any signal above background (data not shown).

8 CXCR2 is another known IL-8 receptor.

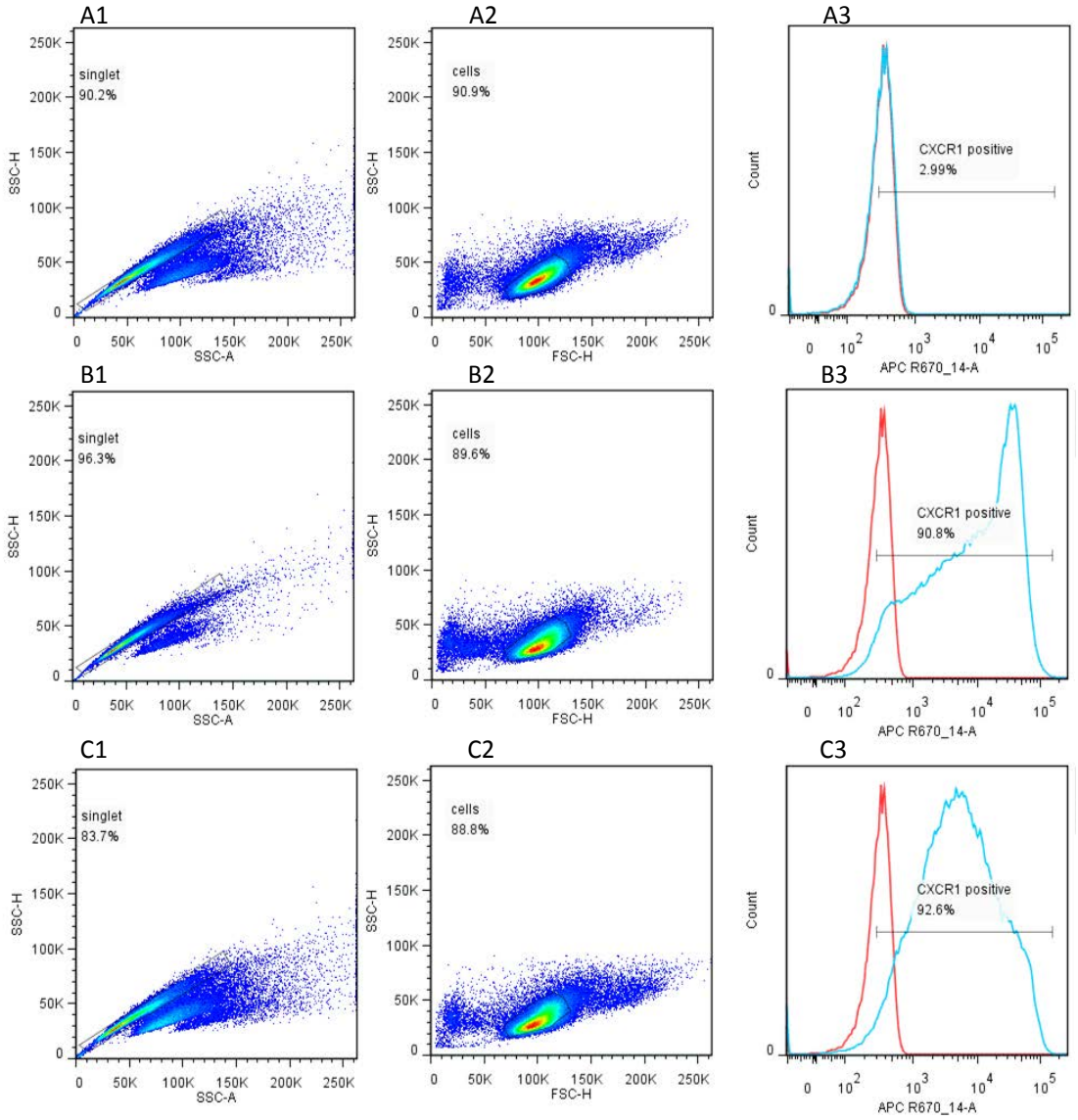


Figure 23 Representative flow cytometry plots for parental and sorted transduced SiHa cells

A1-A3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing parental SiHa cells and histogram of CXCR1 expression by parental SiHa cells, respectively; B1-B3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827G transduced SiHa cells and histogram of CXCR1 expression by CXCR1 827G transduced SiHa cells, respectively; C1-C3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827C transduced SiHa cells and histogram of CXCR1 expression by CXCR1 827C transduced SiHa cells, respectively. By overlaying the image of isotype control on each histogram (red outline), CXCR1 positive cells are identified as the blue peak on the right. Mean fluorescence intensity (MFI) of CXCR1 expression was 357 for parental SiHa cells, 1.7×10^4 for CXCR1 827G transduced SiHa cells and 1.1×10^4 for CXCR1 827C transduced SiHa cells.

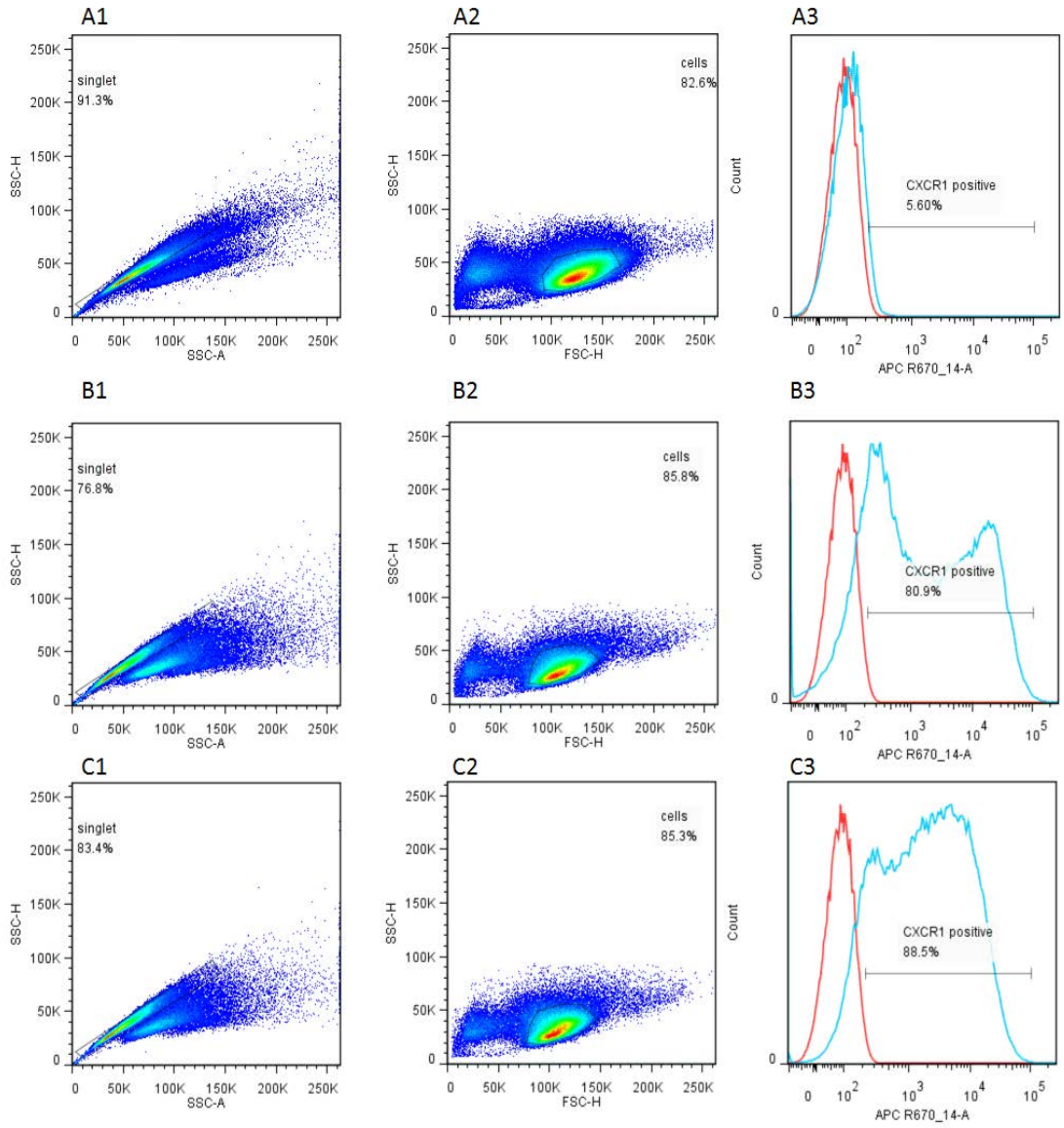


Figure 24 Representative flow cytometry plots for parental and sorted transduced CaSki cells

A1-A3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing parental CaSki cells and histogram of CXCR1 expression by parental CaSki cells, respectively; B1-B3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827G transduced CaSki cells and histogram of CXCR1 expression by CXCR1 827G transduced CaSki cells, respectively; C1-C3: pseudocolor plot with a gate encompassing singlets, pseudocolor plot with a gate encompassing CXCR1 827C transduced CaSki cells and histogram of CXCR1 expression by CXCR1 827C transduced CaSki cells, respectively. By overlaying the image of isotype control on each histogram (red outline), CXCR1 positive cells are identified as the blue peak on the right. Mean fluorescence intensity (MFI) of CXCR1 expression was 286 for parental CaSki cells, 8391 for CXCR1 827G transduced CaSki cells and 5398 for CXCR1 827C transduced CaSki cells.

2.3.6 Assays to investigate functional effects of rs2234671

2.3.6.1 Investigation of the effects of rs2234671 on cell mobilization

2.3.6.1.1 Chemotaxis assay

To assess the functional consequence of the non-synonymous SNP CXCR1 rs2234671 on IL-8 induced cell mobilization, a chemotaxis assay was performed in a transwell system (ABNOVA, 2010) which consists of hanging inserts (8.0μm pore, 6.5mm diameter) and a 24-well cell culture plate. HEK293 cells were seeded at 6.25×10^5/well in a 6-well plate and transiently transfected with CXCR1 expressing plasmid (pCR3-CXCR1 827G or pCR3-CXCR1 827C) or the control vector (pCR3). 24h after transfection, cells were trypsinized with 0.05% Trypsin-EDTA, quenched with 0.05% Soybean Trypsin Inhibitor solution and resuspended in low-serum medium (DMEM with 0.5% FBS) at 1×10^6 cells/ml. The bottom wells of the transwell system were filled with either 600μl DMEM plus 10% FBS medium containing 100ng/ml recombinant human IL-8 (see Figure 25 for the determination of the IL-8 concentration for the chemotaxis assay) or 600μl DMEM plus 10% FBS medium alone. To pre-wet the top compartment, 100μl of low-serum medium (DMEM with 0.5% FBS) was added first into each insert. Then, 100μl of the cell suspension was loaded into the insert and incubated for 6h (37℃ , 5% v/v CO_2, 100% humidity). After incubation, cells on the lower side of the insert (migrated cells) were carefully washed into the bottom well using 1ml PBS. This step was repeated 10 times and all the cell-containing fluid in the bottom well was transferred into a 15ml Falcon tube. Cells were pelleted by centrifugation at 190G for 5min, and then resuspended in 500μl PBS, stained with 5μM

5(6)-Carboxyfluorescein diacetate N-succinimidyl ester (CFSE) in the dark at 37℃ for 10min and quenched with 2ml ice cold DMEM plus 10%FBS medium. After being transferred to a new 24-well plate, cells were settled evenly at the bottom of each well by centrifugation at 190G for 3min and were observed under an EVOS™ digital inverted microscope.

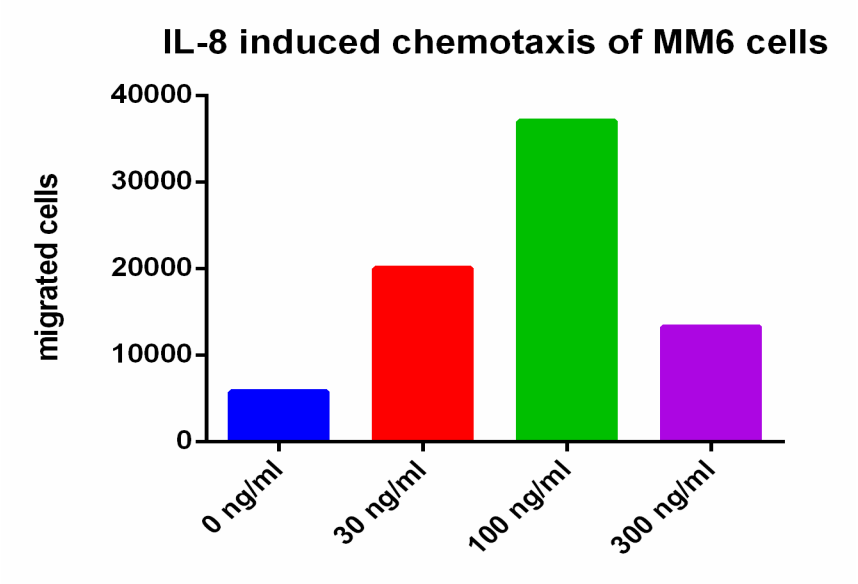


Figure 25 Determination of the IL-8 concentration for the chemotaxis assay

In a single pilot experiment, 1×10^6 MM6 cells, which naturally express CXCR1, was loaded into the insert and incubated in the transwell system for 6h (37°C, 5% v/v CO2, 100% humidity). Cell migration towards different concentration of IL-8 (i.e.0, 30,100 and 300ng/ml) were assessed by counting the total number of migrated cells in the lower chamber of the system after incubation. Since more cells were migrated towards growth medium containing 100ng/ml of IL-8, this concentration was determined as the optimal IL-8 concentration for the following chemotaxis assays. Additionally, 100ng/ml of IL-8 has previously been used in chemotaxis assay for transfected HEK293 cells (Han et al., 2012).

The assay was repeated three times with each condition performed in triplicate. 10 non-overlapping images were taken at 10× magnification for cells migrated from each transwell and the average number of cells on these images was calculated.

Results for each assay were normalized to the average migration of pCR3 transfected cells in response to IL-8 and were presented as chemotactic index (see the equation below). Statistical analysis was performed using ordinary two-way ANOVA test and Sidak's multiple comparisons test.

Chemotactic index=

$$\frac{\text{the number of cells that migrated in each transwell}}{\text{the average number of pCR3 transfected cells that migrated in response to IL}-8\text{ in triplicate tests in each assay}}$$

2.3.6.1.2 Wound healing assay

To assess the functional consequence of the non-synonymous SNP CXCR1 rs2234671 on IL-8 induced tumour cell migration, wound healing assay was performed in 24-well plate with parental CaSki cells and transduced CaSki cells expressing high level of CXCR1. Before seeding the cells, the plate was turned over and drawn with straight reference lines using a waterproof maker pen on the bottom, which cross each row of wells in parallel. Then, each cell line was seeded in sextuplicate at a concentration of 5×10^6cells/well in growth medium containing 10% or 2% of concentration of FBS. Once monolayer was formed in each well, a straight wound was made perpendicularly to the reference line using 200µl tips. The wound was washed twice with PBS and then observed under the EVOS™ digital inverted microscope. Two images were taken above and below the reference line at 10× magnification (see Figure 26 for the detailed experimental design of the wound healing assay) and the initial area of the wound was calculated using the ImageJ software. Briefly, the wound was demarcated manually using the Freehand selections Tool and the area was measured using the Analyze/Measure command.

For each cell line, three wells of cells were then treated with 100ng/ml IL-8 for 12h while the other three were cultured without IL-8. The incubation time was determined by a pilot experiment, in which nearly 50% of the initial area of the wound was covered by migrated CXCR1 827C transduced CaSki cells after 6h of IL-8 treatment (see Figure 27 for the determination of the incubation time for the wound healing assay). Upon completion of the incubation, the wound was washed twice with PBS and two images were taken above and below the reference line at the same locations. The area of the wound after 6h of incubation was also measured using the ImageJ software as described previously. Reduced area for each image was calculated by deducting the final area of the wound from the initial area (unit: pixels square). The migration of the cells was assessed based on the average reduced area of each well. The assay was repeated three times for each condition and statistical analysis was performed using ordinary two-way ANOVA test and Sidak's multiple comparisons test.

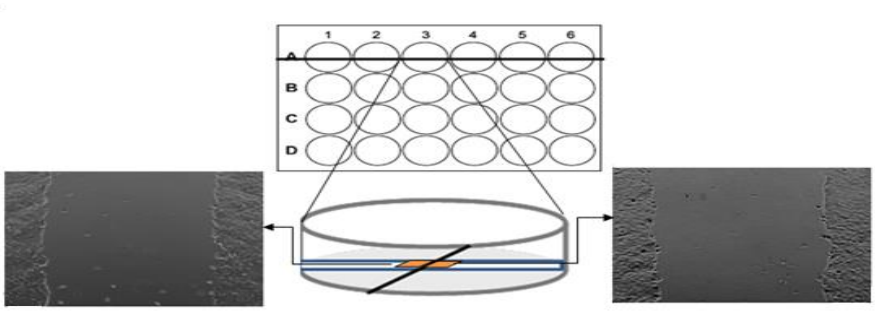


Figure 26 Schematic representation of the experimental design employed in the wound healing assay

Parental CaSki cells and CaSki cells transduced with CXCR1 827G or CXCR1 827C were seeded in 24-well plate and cultured overnight. The next day, a straight scratch was made in each well using 200µl tips perpendicularly to the reference line. Images were obtained at 10X magnification under the EVOS™ digital inverted microscope before and after IL-8 treatment or mock treatment. In the figure above, the thick black line stands for the reference line while blue lines represent boundaries of the wound. The orange areas highlight the sites where images were taken. MRI Wound Healing Tool was obtained from http://dev.mri.cnrs.fr/projects/imagej-macros/wiki/Wound_Healing_Tool. Each image was firstly processed with the Find Edges function of ImageJ (Select Process> Find edges).Then, the "MRI Wound Healing Tool" toolset was selected from the >> button of the ImageJ launcher. Wound area was automatically calculated after pressing the m button. Manual measurements were only accepted if the boundaries could not be properly recognized by the tool.

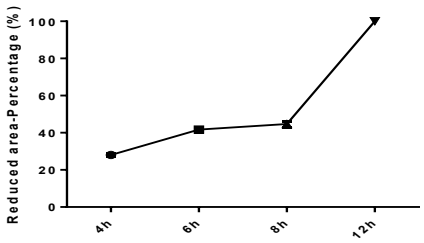


Figure 27 Determination of the incubation time for the wound healing assay

In a pilot experiment, CaSki cells transduced with CXCR1 827C were seeded in 24-well plate and cultured overnight. The next day, a straight scratch was made in each well using 200µl tips perpendicularly to the reference line. Images were obtained at 10X magnification under the EVOS™ digital inverted microscope at 0h, 4h, 6h, 8h and 12h after 100ng/ml IL-8 treatment. Reduced area at each time point for each image was calculated by deducting the final area of the wound from the initial area (unit: pixels square).The time point, at which nearly 50% area of the initial wound was covered by migrated cells, was determined as the optimal observation time point for the following wound healing assays.

2.3.6.2 Investigation of the effects of rs2234671 on cell proliferation

To assess the functional consequence of the non-synonymous SNP CXCR1 rs2234671 on IL-8 induced tumour cell proliferation, proliferation assays were performed with both transduced CaSki cells and SiHa cells expressing high level of CXCR1 using Cell Counting Kit-8. Parental cells and cells transduced with CXCR1 827G or CXCR1 827C were separately seeded in a 96-well plate at a concentration of 5000 cells/well. Then, the cells were treated with a range of concentrations of IL-8 (0, 0.1, 1, 10 and 100ng/ml) for 48h. Each condition was performed in triplicate. Three wells containing only normal growth medium were included in each plate as blank controls. After IL-8 treatment, 10µl of CCK-8 solution was added to each well and incubated at 37°C, 5% CO_2 for 3 hours. Each plate was then centrifuged at 190G for 10 minutes to get rid of bubbles. The absorbance of each well was measured at wavelength of 450nm and 650nm respectively using Synergy HT Multi-Mode microplate Reader. In order to correct for optical imperfections, readings at 650nm was deducted from the readings at 450nm. OD results of all the test wells were normalized to the blank controls within the same plate. Corrected OD value was directly employed in the proliferation rate calculation (see equation below) since it presented a good linear correlation with the cell number in each well. The proliferation assay was repeated three times and statistical analysis was performed using ordinary one-way ANOVA test and Sidak's multiple comparisons test.

Proliferation rate=

$$\frac{\text{Average OD for test wells} - \text{Average OD for correspoding non} - \text{IL} - 8 \text{ wells}}{\text{Average OD for correspoding non} - \text{IL} - 8 \text{ wells}} \times 100\%$$

2.3.6.3 Investigation of the effects of rs2234671 on angiogenesis

To investigate whether CXCR1 rs2234671 affect angiogenesis indirectly by inducing vascular endothelial growth factor A (VEGFA) production, transcript variant 4 of VEGFA (NM_001025368.2, which encodes the most abundant and potent isoform VEGF165) was measured in both transient transfection model and stable transduction model in response to IL-8 stimulation.

2.3.6.3.1 Investigation of the effects of rs2234671 on angiogenesis using transient transfection model

HEK293 cells were transfected with pCR3, pCR3-CXCR1 827G or pCR3-CXCR1 827C vectors as described in **Section 2.3.5.1.2**. Transfected HEK293 cells, together with mock transfected HEK293 cells, were treated with 100ng/ml IL-8 for a range of time periods (i.e. 6h, 8h, 12h and 24h). After treatment, cells were trypsinized, washed with PBS by centrifuging at 190G for 5min and harvested. RNA was isolated from each pellet and then cDNA was made as described in **Section 2.3.1.10**.

The expression of VEGFA transcript variant 4 and GAPDH was analyzed in each cDNA sample using qPCR. VEGFA primers (forward: 5′-TCC TAT CTG AGC CAG ATG ACA TCC-3′; reverse: 5′-CCG GTC TCC CCC GGG TAC C-3′) were designed by Universal ProbeLibrary Assay Design Center according to NM_001025368.2 and used together with Universal ProbeLibrary probe: #51 (Cat#: 04688481001, Roche Diagnostics Ltd. Burgess Hill, UK). qPCR reactions were performed as described in **Section 2.3.1.11**. Table 16 shows the detailed reaction setup.

Table 16 Standard qPCR reaction setup

PCR Component	Volume
TaqMan® Universal PCR Master Mix	*6.5 µl*
PCR-grade H_2O	*5.5µl*
GAPDH Prime& probe Set *Or VEGFA Prime& probe Mix* *(fwd primer + rev primer + probe#51)*	*1µl* *Or(0.45µl +0.45µl +0.1µl)*
cDNA template	*2 µl*
Total reaction volume	*15µl*

Data acquisition and post-acquisition analysis was performed using StepOne™ Software v2.3.VEGFA cycle threshold (CT) value of each sample was normalized to GAPDH by calculating ΔCT (ΔCT= CT_{VEGFA}-CT_{GAPDH}). Then, the relative expression of VEGFA of each sample was calculated using ΔΔCT in comparison with average ΔCT for IL-8 treated mock transfected cells (ΔΔCT=ΔCT $_{\text{test sample}}$ -$\overline{\Delta CT}_{\text{IL-8 treated mock transfected cells}}$). Finally, fold change in each test sample compared with IL-8 treated mock transfected cells was calculated by $2^{-\Delta\Delta CT}$. Each condition was repeated at least three times and the qPCR was performed in duplicate for each sample. Statistical analysis was performed using ordinary two-way ANOVA test and Sidak's multiple comparisons test.

2.3.6.3.2 Investigation of the effects of rs2234671 on angiogenesis using stable transduction models

Parental SiHa cells and SiHa cells transduced with CXCR1 827G or 827C were cultured with or without 100ng/ml IL-8 for 8h and then trypsinized, washed with PBS by centrifuging at 190G for 5min and harvested. RNA was extracted from each pellet using RNeasy Micro Kit and made into cDNA as described in **Section 2.3.6.3.1**. PCR amplification and statistical analysis was also performed as described in **Section 2.3.6.3.1**. VEGFA fold change of each sample was compared with non-IL-8 treated parental SiHa cells. Same procedure was also performed with CaSki and transduced CaSki cells.

Chapter 3
Designing a SNP genotyping array to identify novel SNPs associated with high-risk HPV-related cervical disease

3.1 Introduction

As mentioned in **Chapter 1**, with the results from GWAS studies there is a clear need to bridge the knowledge gap between genotype and phenotype to speed up conversion into real clinical benefits for individual patients. The ideal pipeline would start with the discovery of novel SNPs associated with a particular disease, followed by the replication and validation of the identified genetic associations, and supplemented with the translation of the genetic associations into biological insights.

In comparison with traditional candidate gene approaches, GWA studies have been demonstrated to be a promising approach to identify novel susceptibility loci without needing prior knowledge. Examination of the literature showed that, based on NHGRI-EBI GWAS Catalog (MacArthur et al., 2016), four GWA studies have been published for cervical cancer, together with a newly published GWA study, which identified four potential predictive variants for platinum-based neoadjuvant chemotherapy (Li et al., 2017). However, validation and functional annotation of the significant findings turned out to be very tough.

According to Google Scholar search results, the first GWA study performed in Swedish population for cervical cancer (Chen et al., 2013) has been cited 58 times. This GWA study directed researchers' attention to the HLA region at 6p21.3, within which all 55 SNPs reaching genome-wide significant threshold ($P < 5 \times 10^{-8}$) were identified. The authors of this study further confirmed their findings by independent replication in Swedish population (Chen et al., 2014b) and by pooling the results of two GWAS studies for CIN3 and cervical cancer (Chen et al., 2014a; Chen et al., 2016). Apart from the three independent novel loci discovered in the original GWA study (rs2516448, rs9272143 and rs3117027), the group confirmed the associations

between previously reported HLA loci and risk of CIN3 and cervical cancer, which include HLA-B*07:02, -B*15:01, -DRB1*13:01, -DRB1*15:01,-DQA1*01:03,

-DQB1*06:03 and -DQB1*06:02 and HLA-C*07:02. Also, they identified novel association signals at MICA-A4 and MICA-A5, and further pinpointed potential functional loci, including expression quantitative trait locus (eQTL) rs73730372, cis-eQTL rs9272143 and frameshift mutation rs67841474 (Chen et al., 2016; Chen and Gyllensten, 2014; Chen et al., 2013). A follow-up study was performed in an American population and associations between HLA-B*07:02, -DRB1*13:01, -DQB1*06:03, C*07:02, DQB1*06:04, DRB1*13:02 and cervical adenocarcinoma were identified. However, except for the studies mentioned above, the majority of the 58 publications cited this GWA study only for general background.

The situation was the same for the second most cited GWA study, in which 11 SNPs attained genome-wide significance in combined samples (Shi et al., 2013). Among all the significant SNPs, the association between rs8067378 and cervical cancer was confirmed in a Japanese population and a Polish population, but not in the previously mentioned Swedish population (Chen et al., 2016; Lutkowska et al., 2017; Miura et al., 2016). In the Polish study, the risk allele G was shown to correlate with increased Gasdermin B Expression. The association between rs4282438 and cervical cancer was replicated in a HPV positive Chinese population (Jia et al., 2016). The rest of the associations have neither been confirmed in Chinese population or any other population, and no functionally relevant SNPs have been further identified. Another GWA study performed in a Japanese population did not identify any SNP reaching genome-wide significance and the subsequent studies did not lead to discovery of novel disease related variants (Miura et al., 2014).

There is no doubt that GWA studies contributed to our knowledge about heritability of HPV related cervical disease, but it is not superior to a traditional candidate gene approach for providing well-validated actionable genetic variants. Although reproducible associations have been identified and some novel biological insights have been obtained from published GWA studies, many factors frustrate the continuation of the follow-up studies and functional annotation.

One factor is the conservative threshold set for GWA studies due to statistical reasons. In array-based GWA studies usually around 750,000 tagSNPs are tested in parallel, and correction for multiple testing thus leads to the requirement of p-values of $5x10^{-8}$ for statistical significance. As a consequence, some real but weak or even moderate genetic associations would be missed by GWA studies unless an extremely large sample size was employed. Regarding this point, GWA studies mentioned above serve as good examples. Significant results were limited to within the HLA region at chromosome 6, of which the importance had already been shown extensively in pre-GWAS era. Chen et al showed that these genome-wide significant loci attributed only 2.1% of variation in susceptibility to cervical cancer (Chen et al., 2015), which was much lower than the 27% estimated from nationwide family-based studies (Czene et al., 2002; Magnusson et al., 2000). Even though genome-wide complex trait analysis based on GWAS findings showed that common SNPs (defined as those with GMAF>5%) could tag 88.9% of estimated heritability of cervical cancer, the majority of the disease associated variants could not be unveiled from current GWA studies individually because of the small effect size (Chen et al., 2015). To detect such variants with small individual effects, GWA studies need to be

performed in a much larger sample population. Correspondingly, efforts and costs required will necessarily soar.

Another factor affecting the outcome of GWAS is the coverage of the occurring genetic polymorphisms. GWAS findings highly depend on the genomic coverage of tagSNPs. Research on SNPs has been booming for the last two decades. According to Build 150 (released in 2017), dbSNP now has ~114 million validated SNPs with GMAF<1%, ~6 million SNPs with GMAF between 1% and 5%, and ~12.8 million SNPs with GMAF>5%. Even the number of common SNPs is now nearly twice as many as the one given in the previous estimate (Kruglyak and Nickerson, 2001). Within such a fast-growing field, SNP chips used in published GWASs were very unlikely to have good coverage and efficiency. In fact, whole-genome coverage rates of the chips used in the three published GWA studies of cervical cancer were all shown to be lower than 40% in the corresponding evaluated population (Ha et al., 2014)[9]. Except for those long-established HLA loci, none of the previously reported non-HLA loci associated with cervical cancer was detected in the GWA study performed by Chen et al. Among the 82 cervical cancer-related gene/regions tested, 25 were poorly covered in Human OmniExpress chip (Chen et al., 2015; Johanneson et al., 2014). If selected tagSNPs are not adequately covered, the actual causative SNPs (i.e. in high LD with causative SNPs) association signals would be too weak to be detected.

9 Human OmniExpress had 21% whole-genome coverage rate in evaluated European population. Axiom Genome-Wide CHB1 Array had 37% whole-genome coverage rate in evaluated Asian population. Axiom GW SNP Array 6.0 had 31% whole-genome coverage rate in evaluated Asian population.

In addition, limited access to the raw GWAS data also hinders the progress of functional validation and annotation. Currently, there is no complete GWAS raw dataset available to the public. The two most commonly used databases, GWAS central[10] and GWAS Catalog[11], only host significant results from published GWA studies (Beck et al., 2014; MacArthur et al., 2016). GRASP Search[12] includes supplements and web-based content from publications on the basis of significant findings, but this database offers no more information than a comprehensive manual literature search (Leslie et al., 2014). dbGAP[13] accommodates controlled-access primary data from some studies but it requires a complicated and time-consuming approval process to use (Wong et al., 2017) and unfortunately, this database has few data for HPV-related cervical disease.

Furthermore, even though the use of tagSNPs saves costs for sequencing, it does not provide many functional biological insights. On most occasions, tagSNPs act as surrogate markers rather than actual functional variants. The real causal variants might be found more than 100kb away from the tagSNPs in high LD and they may exist as any forms (e.g. common variant, rare variant, structural variant etc.). (Reich et al., 2001). Therefore, extra efforts, such as extensively re-sequencing and dense-genotyping the up/downstream regions containing disease-associated tagSNPs, are always required to further locate the actual causal SNP. For example, Galarneau and colleagues successfully identified novel independent fetal haemoglobin

10 http://www.gwascentral.org/

11 https://www.ebi.ac.uk/gwas/

12 https://grasp.nhlbi.nih.gov/Search.aspx

13 https://www.ncbi.nlm.nih.gov/gap

associated SNPs and rare functional mutations through fine-mapping (Galarneau et al., 2010). However, because this kind of follow-up methods are usually quite costly and laborious, they have not been widely used and proved to be fruitful. In HR-HPV-related cervical disease, these methods have rarely been used.

Thus, currently causal SNPs accounting for the majority of the heritability of cervical disease are still to be detected. With the rapid reduction in sequencing costs, high expectations have been placed on genome sequencing technologies, such as WGS and whole exome sequencing (WES), to drive this process (Bush and Moore, 2012). These studies are expected to identify novel disease related genetic variations with a higher resolution, to surpass the previously predicted "second generation of GWAS" and fully replace the chip-genotyping. Although some researchers have expressed concerns that the sharply increased amount of data generated from WGS and WES could make the cumbersome statistical analyses even more difficult (King and Nicolae, 2014), the shift from genotyping to genome sequencing now seems inevitable. With the launch of projects like Precision Medicine Initiative and the 100,000 Genomes study (Collins and Varmus, 2015; England, 2016), together with the establishment of data sharing platforms and the development of new computational tools, more novel associations will be identified in the near future and underlying genetic mechanisms for HPV-related cervical cancer will be uncovered.

While awaiting the progress from those large-scale genotyping or sequencing projects, researchers like me are also keen to make efforts to bridge the knowledge gap between genotypes and the underlying biological mechanisms. In the absence of strong financial and information resources, we still have to rely on traditional candidate gene approaches to identify novel disease associated variants and literature

mining to select potentially functional variants within candidate genes. The rationale of this method is pleiotropy (Stearns, 2010). Pleiotropy in this context is to describe a phenomenon that certain genes or some SNPs within or close to those genes are associated with multiple phenotypes and that these phenotypes are not necessarily overlapping or related to each other. A classic study published in 2011 confirmed the commonness of pleiotropy based on GWAS findings and showed the characteristics of pleiotropic SNPs. In this study, ~17% and ~5% known disease associated genes and SNPs respectively were conservatively estimated to be pleiotropic. Compared to non-pleiotropic SNPs, pleiotropic SNPs were shown to be enriched in exonic regions and more likely to be structurally functional (Sivakumaran et al., 2011). The ubiquity of pleiotropy was further supported by a number of studies (Bulik-Sullivan et al., 2015; Pickrell et al., 2016). According to the findings from these studies, pleiotropic SNPs are plausible candidates to start a new candidate gene based association study. They have association signals detected previously in multiple diseases only because they are causal SNPs themselves or closely correlate with the real causal SNPs. If they are truly causal, they may affect multiple traits through a shared biological process or though different pathways, which makes them conspicuous targets for further functional assays. Although often being criticized for low region coverage, the candidate gene approach benefits from its flexibility in choosing potential targets, a characteristic welcomed by researchers aiming to make functional translations. When performing such an association study, researchers can focus on pleiotropic SNPs with known functions or those that reside within the coding regions or regulatory regions of a certain gene. If selected pleiotropic SNPs are known to be functional and they are associated with HPV-related cervical disease, the knowledge

about their functional effect can directly be employed to further elucidate the pathogenesis. On the other hand, if selected pleiotropic SNPs with no functional annotations are shown to be associated with HPV-related cervical disease, once their functional effects are identified in functional assays, the findings might be generalized to other related traits. The whole process, to some extent, is similar to drug repurposing. Since this process effectively reduces costs and saves efforts, it was chosen for this project.

The aim of this chapter is to describe the panel design for the subsequent SNP array study. The panel targeted identification of novel functional variants in HPV-related cervical disease. The selection of the SNPs for this study was based on the hypothesis that **"some known pleiotropic SNPs within genes affecting HR-HPV-related diseases are functionally associated with HR-HPV-related diseases"**. All candidate SNPs were selected from pleiotropic SNPs within a group of genes of interest on the basis of a comprehensive literature search. The selection of the candidate genes is described in **Section 3.2**.The literature search performed for the extraction of pleiotropic SNPs is described in **Section 3.3**. And the final panel applied in the SNP array is described in **Section 3.4**.

3.2 Selection of the candidate genes for the SNP array

For ease of handling, I mainly focused on a small group of genes/regions rationally selected from a chemokine profile study of cervical disease previously performed in our group, and, from two published array-based genetic association studies (Safaeian et al., 2012; Wang et al., 2010). The selected panel consisted mainly of genes involved in immune responses, DNA repair, oncogenesis and viral infection and cell entry.

3.2.1 Candidate genes from the chemokine profile study

The chemokine profile study was previously performed with R&D Systems Proteome Profiler™ Human Chemokine Array using protein samples extracted from LBC samples. A number of chemokines were found to be up-regulated in pooled HR-HPV positive CIN 1, CIN2 or CIN3 samples, compared with pooled HR-HPV negative samples. Since very few SNPs within the chemokine or chemokine receptor genes have been investigated in HPV-related diseases (as mentioned in Chapter 1), genes encoding these chemokines were all chosen as candidate genes for further selection steps. In particular, some chemokines (e.g.CCL2, CXCL1, CXCL8, CXCL10, CXCL12 etc.) were found to be up-regulated in HR-HPV positive high grade cervical disease on both the Luminex® and the microarray platforms. Thus SNPs within genes encoding these chemokines or their corresponding receptors had some priority over other SNPs in the final selection step.

3.2.2 Candidate genes from two published array-based genetic association studies

Prior to the GWAS era, two consecutive large-scale array-based association studies performed in the same Costa Rican cohort aroused widespread attention (Safaeian et al., 2012; Wang et al., 2010). By extensively genotyping SNPs within rationally selected genes (e.g. immune genes, DNA repair genes, viral binding and cell entry genes, etc.), a number of novel susceptibility loci were identified for persistent HPV infection and for the progression to pre-cancer/cancer (See Chapter 1 for details). Some of the findings were shown to be reproducible in studies involving simple replication (Colacino-Silva et al., 2017; Famooto et al., 2013). In order to increase

the power to explore the underling genetic architecture and disease mechanisms, the adaptive rank truncated product method was used to condense SNP level p values within genes and regions in these two array-based studies. Briefly, top K smallest single SNP p values within a certain gene/region were combined through the truncated product method and the optimal truncation point k was determined in an adaptive way (Dudbridge and Koeleman, 2004; Yu et al., 2009).

In complex diseases, it is very common to find that a given gene contains more than one causal SNP and the association signals generated by these causal SNPs are loosely captured by other SNPs in strong or moderate LD. Since tagSNPs selected in these two studies could not completely cover the gene/regions, the gene-level summary statistics sparked later researchers' interest in mining for novel disease associated variants either dependent or independent of the original association signals within the same related gene/regions. For example, Castro and colleagues further genotyped 12 novel SNPs within the TMC6 and TMC8 genes and successfully identified two novel SNPs (rs16970849 and rs2290907) independently associated with cervical cancer susceptibility in the Swedish population (Castro et al., 2012). In this project, I included all 37 genes with gene-level p value <0.01 in the two mentioned array-based studies as candidate genes but mainly focused on pleiotropic SNPs that are independent of the original 19 significant findings. Since only a small fraction of women with persistent HPV infection will finally develop pre-cancer lesions and cervical cancer, genes involved in this transition merit more attention in further investigations. For this reason, 17 genes with p value <0.05 in tests for associations with progression to cervical pre-cancer/cancer were also

included. However, SNPs within genes with smaller p values were given higher priority in the final selection step.

3.2.3 The final list of candidate genes

As mentioned above, candidate genes were confined to those that were highlighted in HPV-related cervical disease and with key roles in multiple biological processes. The final candidate gene list consists of 22 chemokine genes, 7 chemokine receptor genes and 54 other genes that are potentially associated with persistent HPV infection or cervical cancer (see Table 17 for the complete list of candidate genes/regions involved in the project). Borrowing ideas from the pathway-based approach, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)[14] was also used to provide supplementary information for the next selection step at pathway level (see Appendix 2 for the protein-protein interactions among the proteins encoded by candidate genes). However, since no distinct pathways other than chemokine system related processes were identified, this piece of information was not taken into account when determining the priority of the candidate SNPs in the next step.

14 STRING: functional protein association networks (https://string-db.org/)

Table 17 Candidate genes/regions involved in this project

Candidate gene list (n=83)				
Gene symbol	Gene name	Gene ID	Source	Note
CCL1	chemokine (C-C motif) ligand 1	6346	chemokine profile study	-
CCL2	chemokine (C-C motif) ligand 2	6347	chemokine profile study	-
CCL3	chemokine (C-C motif) ligand 3	6348	chemokine profile study	-
CCL5	chemokine (C-C motif) ligand 5	6352	chemokine profile study	-
CCL7	chemokine (C-C motif) ligand 7	6354	chemokine profile study	-
CCL17	chemokine (C-C motif) ligand 17	6361	chemokine profile study	-
CCL18	chemokine (C-C motif) ligand 18	6362	chemokine profile study	-
CCL19	chemokine (C-C motif) ligand 19	6363	chemokine profile study	-
CCL21	chemokine (C-C motif) ligand 21	6366	chemokine profile study	-
CCL22	chemokine (C-C motif) ligand 22	6367	chemokine profile study	-
CCL26	chemokine (C-C motif) ligand 26	10344	chemokine profile study	-
CCL28	chemokine (C-C motif) ligand 28	56477	chemokine profile study	-
CXCL1	chemokine (C-X-C motif) ligand 1	2919	chemokine profile study	-
CXCL4	chemokine (C-X-C motif) ligand 4	5196	chemokine profile study	-
CXCL5	chemokine (C-X-C motif) ligand 5	6374	chemokine profile study	-
CXCL8	chemokine (C-X-C motif) ligand 8	3576	chemokine profile study	-
CXCL9	chemokine (C-X-C motif) ligand 9	4283	chemokine profile study	-
CXCL10	chemokine (C-X-C motif) ligand 10	3627	chemokine profile study	-
CXCL11	chemokine (C-X-C motif) ligand 11	6373	chemokine profile study	-
CXCL12	chemokine (C-X-C motif) ligand 12	6387	chemokine profile study	-
IL16	interleukin 16	3603	chemokine profile study	-
MDK	midkine	4192	chemokine profile study	-
CCR2	chemokine (C-C motif) receptor 2	729230	chemokine profile study	-
CCR10	chemokine (C-C motif) receptor 10	2826	chemokine profile study	-
CXCR2	chemokine (C-X-C motif) receptor 2	3579	chemokine profile study	-
CXCR1	chemokine (C-X-C motif) receptor 1	3577	chemokine profile study	-
XCR1	chemokine (C motif) receptor 1	2829	chemokine profile study	-
CCR1	chemokine (C-C motif) receptor 1	1230	chemokine profile study	-
DARC	atypical chemokine receptor 1 (Duffy blood group)	2532	chemokine profile study	-
ALOX5	arachidonate 5-lipoxygenase	240	published array	P<0.01
ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	19	published array	P<0.01
CP	ceruloplasmin (ferroxidase)	1356	published array	P<0.01
C1QA	complement component 1, q subcomponent, A chain	712	published array	P<0.01
C1QR1	CD93 molecule	22918	published array	P<0.01
C1RL	complement component 1, r subcomponent-like	51279	published array	P<0.01
DDX1	DEAD (Asp-Glu-Ala-Asp) box helicase 1	1653	published array	P<0.01
DMC1	DNA meiotic recombinase 1	11144	published array	P<0.01
DUT	deoxyuridine triphosphatase	1854	published array	P<0.01
ENPP1	ectonucleotide pyrophosphatase/phosphodiesterase 1	5167	published array	P<0.01
EVPL	envoplakin	2125	published array	P<0.01
GC	group-specific component (vitamin D binding protein)	2638	published array	P<0.01
GDF10	growth differentiation factor 10	2662	published array	P<0.01

Candidate genes/regions involved in this project (continued)

Candidate gene list (n=83)				
Gene symbol	Gene name	Gene ID	Source	Note
GDF2	growth differentiation factor 2	2658	published array	P<0.01
GJA1	gap junction protein, alpha 1, 43kDa	2697	published array	P<0.01
GSTCD	glutathione S-transferase, C-terminal domain containing	79807	published array	P<0.01
GTF2H4	general transcription factor IIH, polypeptide 4, 52kDa	2968	published array	P<0.01
IFNG	interferon, gamma	3458	published array	P<0.01
IL2RA	interleukin 2 receptor, alpha	3559	published array	P<0.01
ILDR1	immunoglobulin-like domain containing receptor 1	286676	published array	P<0.01
LY96	lymphocyte antigen 96	23643	published array	P<0.01
MBD3L1	methyl-CpG binding domain protein 3-like 1	85509	published array	P<0.01
MICA	MHC class I polypeptide-related sequence A	100507436	published array	P<0.01
MOCS2	molybdenum cofactor synthesis 2	4338	published array	P<0.01
NEIL2	nei endonuclease VIII-like 2 (E. coli)	252969	published array	P<0.01
OAS3	2'-5'-oligoadenylate synthetase 3, 100kDa	4940	published array	P<0.01
PGR	progesterone receptor	5241	published array	P<0.01
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subuni	5290	published array	P<0.01
PRDX3	peroxiredoxin 3	10935	published array	P<0.01
PRLH	prolactin releasing hormone	51052	published array	P<0.01
RPS19	ribosomal protein S19	6223	published array	P<0.01
SULF1	sulfatase 1	23213	published array	P<0.01
TELO2	telomere maintenance 2	9894	published array	P<0.01
MED24	mediator complex subunit 24	9862	published array	P<0.01
TMC6	transmembrane channel-like 6	11322	published array	P<0.01
TMC8	transmembrane channel-like 8	147138	published array	P<0.01
TYMS	thymidylate synthetase	7298	published array	P<0.01
CADM1	cell adhesion molecule 1	23705	published array	P<0.05
CD83	CD83 molecule	9308	published array	P<0.05
CDKN2B	cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)	1030	published array	P<0.05
DCLRE1A	DNA cross-link repair 1A	9937	published array	P<0.05
FANCD2	Fanconi anemia, complementation group D2	2177	published array	P<0.05
FASLG	Fas ligand (TNF superfamily, member 6)	356	published array	P<0.05
FLJ35220	endonuclease V	284131	published array	P<0.05
ITGA4	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 rece	3676	published array	P<0.05
ITGA7	integrin, alpha 7	3679	published array	P<0.05
LIG3	DNA ligase 3	3980	published array	P<0.05
OAS1	2'-5'-oligoadenylate synthetase 1, 40/46kDa	4938	published array	P<0.05
OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	4939	published array	P<0.05
POLN	polymerase (DNA directed) nu	353497	published array	P<0.05
TNF	tumor necrosis factor	7124	published array	P<0.05
TOPBP1	topoisomerase (DNA) II binding protein 1	11073	published array	P<0.05
WIF1	WNT inhibitory factor 1	11197	published array	P<0.05
WRN	Werner syndrome, RecQ helicase-like	7486	published array	P<0.05

3.3 Selection of pleiotropic SNPs within candidate genes

Literature mining was used for the selection of candidate SNPs within candidate genes/regions (see Figure 28 for the flow chart showing the whole selection procedure for candidate SNPs). In order to pick out all previously reported pleiotropic SNPs, I performed a comprehensive literature search using PubMed SNP database on all the candidate genes/regions (https://www.ncbi.nlm.nih.gov/pubmed/) and carefully reviewed all the SNPs that have been investigated in articles published in English before February 2015 within each gene/region.

Since an extremely large sample size would have been required to detect rare variant associations but only a limited number of clinical samples were available for the SNP array, SNPs with a GMAF<5% were excluded in the first round of selection. 70 of the 83 candidate genes had at least one published common SNP (GMAF≥5%) in the PubMed SNP database. No published common SNPs were identified for the other 13 candidate genes, including CXCL4, Midkine, EVPL, GJA1, MBD3L1, MOCS2, PRLH, FLJ35220, ITGA7, GSTCD, XCR1, MED24 and C1RL.

In total, 630 validated common SNPs were identified in candidate genes/regions. Among all the validated common SNPs (see Appendix 3 for the complete list), only clinically significant SNPs or disease associated SNPs were included in the second round of selection. 275 of these 630 SNPs were found to be associated with at least one disease. In order to fit in a custom multiplex SNP array with a reasonable cost, a set of 32 SNPs was further selected from the 275 disease-related SNPs based on at least one of the following criteria: 1) previously shown to be functional; 2) non-synonymous SNPs or SNPs reside within regulatory regions (e.g. promoter,

enhancer, insulator, microRNA binding site etc.); 3) in high linkage disequilibrium with other candidate SNPs; 4) within chemokine or chemokine receptor genes identified in the previous chemokine profile study; 5) previously shown to be associated with multiple diseases, especially infectious diseases and cancers; 6) with high GMAF; 7) within genes with confirmed functional effects in HPV associated diseases.

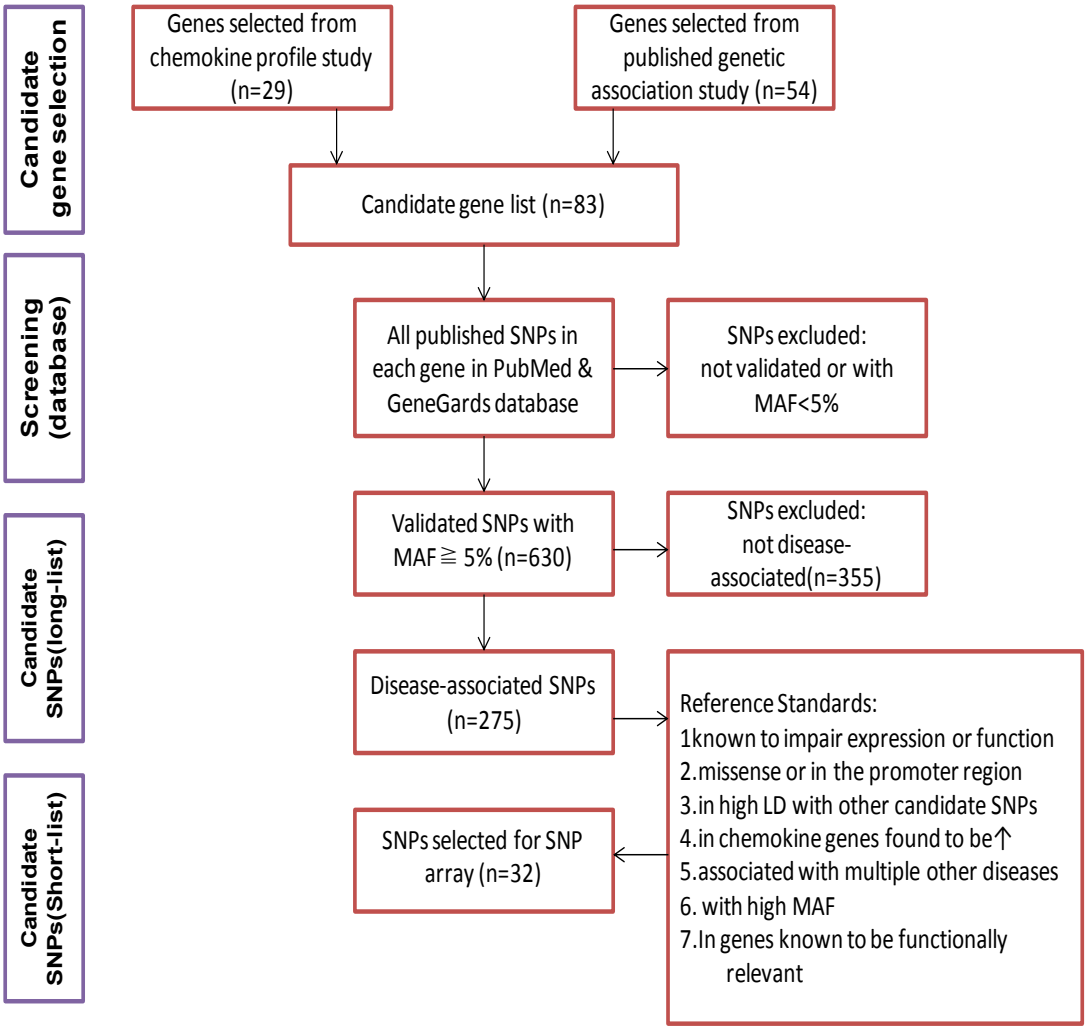


Figure 28 Flow chart with the selection procedure for candidate SNPs

3.4 The final panel of 32 SNPs involved in the SNP array

A subset of 32 SNPs was chosen for the final SNP array (see Table 18 for the full list). These SNPs were located within 22 of the 83 candidate genes, including 7

chemokine genes found to be up-regulated in HR-HPV positive high grade cervical disease in the chemokine profile study, 3 chemokine receptor genes which encode receptors bound for more than one of the chemokines of interest, and 12 genes (9 with $p<0.01$ and 3 with $p<0.05$) from the two published array-based genetic association studies. All of the 32 SNPs are biallelic. The global minor allele frequency of these 32 SNPs ranges from 0.06 to 0.48. Most of these 32 SNPs were in high LD with other candidate SNPs in the long-list and thereby were also used as proxies for other unselected SNPs. Among the 32 SNPs, 17 were shown to have impacts on gene expression or function; 31 were involved in other cancers or infectious diseases; 4 were missense SNPs and 13 were found in the promoter region. In addition, 29 of the 32 SNPs had off-the-shelf TaqMan® SNP Genotyping Assays available. Custom TaqMan® Assays were designed by Applied Biosystems™ for the other 3 SNPs, including rs16970849, rs2234671 and rs17849071. Another noteworthy point is that TMC8 rs16970849, CD83 rs853360, TNF-α rs1800629, TNF-α rs361525 were deliberately included in the array as they all had been reported to contribute to cervical cancer susceptibility in other populations (Bodelon et al., 2012; Castro et al., 2012; Liu et al., 2012a; Zhang et al., 2007). Their associations with HPV-related diseases would be assessed in Scottish population in this project. Additionally, OAS1 rs34137742 had been shown to be associated with cervical cancer in a pilot experiment in our group compared with female European controls extracted from 1000Genomes database (data not shown). This association would be re-validated in comparison with samples from HPV negative cytologically normal Scottish women as controls.

Table 18 the 32 candidate SNPs selected for the SNP array

gene symbol	SNPs	SNPs location	Allele	MAF	Disease/phenotype	PMID	cancer or infectious disease-related	Functionally relevant	upstream or missenese
CCL2	rs2857656	upstream variant 2KB	C/G	C=0.4209/2108	Tuberculosis, Carotid atherosclerosis, Non contact soft tissue injuries	18940815 19506371 23890452	YES	No	upstream
CCL2	rs3760396	upstream variant 2KB	C/G	C=0.0940/470	Non-small-cell lung carcinoma, Rosiglitazone or pioglitazone treatment, ovarian cancer	21514686 18996102 25289731	YES	No	upstream
CCL5	rs2107538	upstream variant 2KB	C/T	T=0.3077/1541	Papillary thyroid cancer, Post-transplantational diabetes mellitus	23957698 20805685	YES	No	upstream
CXCL1	rs4074	intron variant	A/G	G=0.4044/2024	Hepatitis C	22173151	YES	YES	
CXCL1	rs3117604	upstream variant 2KB	C/T	T=0.4728/2367	Ischemic stroke	23198952		No	upstream
CXCL8	rs4073	upstream variant 2KB	A/T	T=0.4778/2392	Colorectal cancer, Sepsis, Myocardial infarction, Non-small-cell lung carcinoma, Oral cancer, Apical periodontitis, Idiopathic pulmonary fibrosis, Childhood IgA nephropathy, Oral lichen planus, Non-Hodgkin lymphoma	25139485 25000179 24462138 23831257 23545310 22788685 21649933 21214373 19842025 19066394	YES	YES	upstream

gene symbol	SNPs	SNPs location	Allele	MAF	Disease/phenotype	PMID	cancer or infectious disease-related	Functionally relevant	upstream or missenese
CXCL10	rs3921	intron variant, utr variant 3 prime	C/G	C=0.3075/1540	liver fibrosis in HCV/HIV coinfection	25559603	YES	No	
CXCL12	rs1801157	utr variant 3 prime	A/G	T=0.1855/928	Multiple myeloma, Acute lymphocytic leukemia , Breastcancer, Esophagogastric cancer, Non-Hodgkin lymphoma, Myocardial infarction, HIV, Hepatocellular carcinoma, Premature ovarian failure	23711392 23653000 21643956 19927352 19821058 19601773 19327121 21296802	YES	YES	
IL16	rs11556218	missense, nc transcript variant	G/T	G=0.1464/733	Nasopharyngeal carcinoma, Coronary artery disease, Prostate cancer, Systemic lupus erythematosus, Colorectal cancer, glioma, COPD	24101193 21703255 22923025 19298795 19073878 25166752 24138069	YES	YES	missense
IL2RA	rs7072793	upstream variant 2KB	C/T	C=0.4511/2258	Breast cancer, Type 1 Diabetes mellitus	22213266 19956099	YES	YES	upstream
IL2RA	rs2104286	intron variant	A/G	C=0.1306/654	Bladder cancer, Multiple sclerosis, Rheumatic arthritis, Giant cell arteritis	24931268 24770783 23529819 20810507	YES	YES	

gene symbol	SNPs	SNPs location	Allele	MAF	Disease/phenotype	PMID	cancer or infectious disease-related	Functionally relevant	upstream or missenese
MICA	rs2596542	intron variant, upstream variant 2KB	A/G	T=0.4215/2111	Hepatic cell carcinoma, Hepatitis C, Hepatitis B	24357186 24023482 24010643	YES	YES	upstream
NEIL2	rs804270	intron variant, utr variant 5 prime	C/G	G=0.4820/2414	Squamous cell carcinoma of the oral cavity and oropharynx	18594018	YES	No	
OAS1	rs10774671	intron variant, splice acceptor variant, utr variant 3 prime	A/G	G=0.3856/1930	Hand-foot-and-mouth disease, Hepatitis C, Hepatitis B, West Nile virusinfection, Type 1 Diabetes mellitus	25059424 22710942 19799013 19247438 16014697	YES	YES	
PGR	rs1042838	intron variant, missense, nc transcript variant	G/T	A=0.0715/357	Endometrial cancer, Recurrent pregnancy loss , Migraine-associated vertigo, Breast cancer	22633539 21086036 17609999 15632380	YES	No	missense
PGR	rs11224561	nc transcript variant, utr variant 3 prime	C/T	T=0.3013/1509	endometrial cancer	21148628	YES		
PIK3CA	rs2699887	intron variant, upstream variant 2KB	A/G	T=0.1176/589	Non-small-cell lung carcinoma, Endometrial cancer	24077347 22146979	YES	No	upstream
PIK3CA	rs17849071	intron variant	G/T	G=0.1026/514	Breast cancer	24908061	YES	YES	

gene symbol	SNPs	SNPs location	Allele	MAF	Disease/phenotype	PMID	cancer or infectious disease-related	Functionally relevant	upstream or missenese
SULF1	rs2623047	upstream variant 2KB	C/T	A=0.4736/2372	Breast cancer, Ovarian cancer	24911625 21214932	YES	YES	upstream
TMC8	rs12452890	missense, nc transcript variant, synonymous codon	A/G	G=0.4597/2302	Squamous cell skin	24913986	YES	No	missense
TMC8	rs16970849	downstream variant 500B, intron variant	A/G	A=0.1430/715	Cervical diseases	21387292	YES	No	downstream
TYMS	rs2790	intron variant, utr variant 3 prime	A/G	G=0.2939/1472	Non-small-cell lung carcinoma, Gastric cancer	24997136 24756984	YES	Yes	
CD83	rs853360	intron variant	C/T	T=0.2354/1178	Cervical diseases	22134374 18056445	YES	YES	
TNF	rs361525	upstream variant 2KB	A/G	A=0.0609/305	Inflammatory bowel disease, End-stage liver disease, Psoriasis, Cervical cancer, Rheumatic arthritis, Hepatic cell carcinoma, Chronic obstructive pulmonary disease, Graves disease], Breast cancer, Juvenile dermatomyositis, Asthma	24776844 24361409 24252077 21670964 21385363 21080879 20299531 19732761 19423537 19035492 17450233	YES	YES	upstream
TNF	rs1799724	downstream variant 500B, upstream variant 2KB	C/T	T=0.0990/495	Polycystic ovary syndrome, Achalasia, Lung cancer, Bone mineral density, Inflammatory bowel disease	25083576 24259423 24139238 18551993 12019209	YES	YES	upstream

gene symbol	SNPs	SNPs location	Allele	MAF	Disease/phenotype	PMID	cancer or infectious disease-related	Functionally relevant	upstream or missenese
TNF	rs1800629	upstream variant 2KB	A/G	A=0.0903/452	Tuberculosis, Thyroid disease, Complicated skin and skin structure infections, Sepsis, Asthma, Periodontitis, Psoriasis, Gastric cancer, Prostate cancer, Acute kidney injury, Breast cancer, Non-Hodgkin's lymphoma, Sjogren's syndrome, Cervical diseases, Hepatitis B, Chronic obstructive pulmonary disease, Lung cancer, RA, Schizophrenia & Bipolar affective disorder	25239251 25127106 25022448 25000179 24936650 24905365 24252077 24142527 24037955 23796916 23530106 22649007 22294627 21670964 19933216 19773451 18713756 18515978	YES	YES	upstream
CXCR2	rs1126579	utr variant 3 prime	C/T	T=0.3968/1986	Bile duct cancer& biliary stones, Bladder cancer, Lung cancer	18676870 19252927 25480945	YES	YES	
ACKR2	rs4683336	intron variant	C/T	C=0.3664/1835	Hepatitis C	18822328	YES	YES	
CXCR1	rs2234671	missense	C/G	G=0.1442/721	Hepatitis B	23396733	YES	No	missense
CXCL12	rs266085	intron variant	C/T	T=0.3738/1872	immune thrombocytopenia, cervical carcinoma	23078136 19788587	yes	No	
OAS1	rs34137742	intron variant	C/T	T=0.1030/516	West Nile virus prostate cancer	21935451 21638280	yes	no	
IFNG	rs2069705	upstream variant 2KB	C/T	G=0.4778/2392	breast cancer, SLE	20418110 19919944	yes	No	upstream

3.5 Discussion

Very few follow-up studies have been performed to verify the results of the previous association studies and to further locate functional variants within the genes differentially expressed in HPV-related diseases. According to past experience, traditional candidate gene/SNP studies based solely on prior knowledge and hypothesis are quite inefficient and unlikely to accelerate the discovery of functional variants. Simply deep re-sequencing the genomic regions highlighted by GWASs or array-based studies for causal variants is costly and was unfeasible in the current project. There are no conventional guidelines or well-established selection procedures available to cost-effectively follow up existing genetic association studies and pinpoint SNPs for functional studies. Hence, an evidence-based strategy was proposed for choosing potentially functional SNP on the basis of the biological pleiotropy at the allelic level, with the goal of improving the power to detect novel causal variants and minimizing the number of samples needed. Since this strategy allowed the prioritization of known functional SNPs, non-synonymous SNPs and SNPs in promoter regions, it was expected to speed functional variant identification in HPV-related diseases in a small scale project with a limited budget. It might also be applied in other diseases for the identification of functional SNPs

This whole project was started in October, 2014 and the literature search was completed by the end of January, 2015. Considering present-day standards, several limitations of the panel design have to be acknowledged. Unavoidably, this design suffered from a common failing of many candidate gene approach based studies: inadequate coverage of the candidate gene/regions. Although rationally selected SNPs rather than randomly chosen SNPs were employed to improve the

power to detect the potentially functional variants and reduce the number of samples required, the possibility that real causal variants were missed within the same gene/region could not be excluded. Furthermore, since only one or two SNPs in weak linkage disequilibrium or in complete linkage equilibrium within each gene were included in the SNP array, it was impossible to perform haplotype based analysis or assess the joint effects of the association signals.

Another limitation is the lack of knowledge regarding genomic characterization and gene annotation. Identification of disease related SNPs is a very active and dynamic research field. A lot of powerful predictive tools used extensively at the moment were either not available or not being aware at the beginning of the project. For example, to determine priority of the candidate genes or SNPs at the gene-level, a piece of evidence frequently used now is the survival assessment of each gene based upon TCGA cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) database[15]. Using the newly released web based tool GEPIA (Tang et al., 2017), it is really convenient to evaluate the prognostic value of a particular gene in the mentioned database. Out of all the chemokines and chemokine receptors identified to date, higher expression of CCL20 [hazard ratio (HR) 1.6, P = 0.043], CXCL1 [HR1.7, P = 0.035], CXCL3 [HR1.8, P = 0.019], CXCL4 [HR1.7, P = 0.029], CXCL5 [HR1.8, P = 0.012], CXCL8 [HR2.9, P = 3.4e-05] and CXCR1 [HR1.6, P = 0.048] were predictive for poor overall survival in this dataset. While higher expression of CCL17 [HR0.49, P = 0.003], CCL25 [HR0.6, P = 0.035], CX_3CL1 [HR0.48, P = 0.0036], CCR2 [HR0.52, P = 0.0067] and CCR7 [HR0.47, P = 0.0019] was associated with a reduced overall survival in cervical cancer patients.

15 https://portal.gdc.cancer.gov/projects/TCGA-CESC

Among all the candidate genes selected from the two published array-based studies, GJA1 [HR1.8, P = 0.02], TMC8 CCR7 [HR0.46, P = 0.0014], TYMS [HR0.62, P = 0.044] and CD83 [HR0.61, P = 0.041] were also found to be associated with the overall survival (see Appendix 4 for the Kaplan Meier curve of each gene mentioned). Except for the predictive tools at gene level, there were also some tools used to forecast the functional effects of known SNPs, such as SIFT[16], PolyPhen-2[17], SNPnexus[18] and MutationTaster[19]. If findings like these were available or aware at the time of panel design, unquestionably they would affect the strategy making. However, although the current panel failed to highlight all the above genes and SNPs within them, it still successfully covered some intriguing targets. For instance, at least one SNP was selected in the actual array for CXCL1, CXCL8, CXCR1, TMC8, TYMS and CD83 and most of them were known as functional variants or located within functional regions of the corresponding genes.

Furthermore, my newness to this research field and my inexperience at the beginning of the project also contributed to the limitations of the panel design. As a rookie researcher, I was incapable of selecting disease related SNPs with a critical eye. The initial inclusion of the disease related SNPs only relied on the statistical significance reported in the main text or supplementary documents but without any further evaluation. It now seems that more attention should have been paid to the details of the studies since mistakes were not infrequently observed in published association

16 http://sift.jcvi.org/

17 http://genetics.bwh.harvard.edu/pph2/

18 http://snp-nexus.org/

19 http://www.mutationtaster.org/

studies. For example, improper interpretation of the statistical analysis was observed in a study mentioned previously (Jia et al., 2016). In their table 3, illogical comparisons were made and conclusions without biological meaning were drawn. Similar mistake was also observed in the supplementary Table 3 of another association study (Hsing et al., 2008). Additionally, overt plagiarism and potentially fabrication of data was also found in one publication (Li et al., 2016)[20]. Taking all these conditions into account and re-evaluating the whole list of published SNPs, the current long list of candidate SNPs might be slightly changed. However, since more stringent standards were applied in the following selection steps, candidate SNPs in the short list were unlikely to be affected.

The rationally selected panel used was an active attempt to pinpoint potentially functional variants in HPV-related disease in a rational and inexpensive way. Although it was not perfect and only fitted in a small scale array, its feasibility and effectiveness are shown in the following results chapters. Just as statesman Deng Xiaoping said "It does not matter if a cat is black or white, so long as it catches mice". Like everyone working in the same field, I look forward to every new cost-saving method and strategy that could accelerate the identification of functional variants and I believe a widely applicable method will come out in the near future, along with the rapid development of sequencing techniques. But still, no matter how fancy and comprehensive are the in silico prediction tools and computational methods, the key results and the basis of clinical translation would always be direct functional and biological evidence.

20 This publication plagiarized a previous study (PMID: 25113639)

Chapter 4
Results of the SNP array and further validation of the significant SNP in a broad spectrum of HPV-related diseases

4.1 Introduction

As mentioned before, a few SNPs associated with HPV-related diseases, especially cervical cancer, were identified through traditional candidate gene based association studies and GWA studies. These findings provide unequivocal evidence that cervical cancer, like other complex diseases, is highly polygenic and the majority of disease related SNPs have fairly small effects, of which the odds ratios are usually below 2 (Wang et al., 2015).

Most of these studies were performed with a case-control setting. As a classic experimental design, the history of case control study can be traced back to 1926 (Paneth et al., 2002). It was not only the preferred design for genetic association studies but was also extensively used for other epidemiological studies due to its effectiveness and resource saving property. In this project, the SNP array was also performed in a case-control setting. The association between the presence of a particular SNP and disease susceptibility was assessed by comparing the differences of allele frequencies or alternatively genotype frequencies between case and controls.

Theoretically, genomic DNA extracted from any source can be used for genetic association studies, with the success and outcomes of such studies being heavily reliant on the genotyping or sequencing methods employed. The most commonly used sample types including blood, saliva, normal cytology specimens and tissue specimens. In recent years, a consensus has been gradually reached in the cancer research field that SNPs can be distinguished from somatic mutations through their concordance between germline DNA and tumour DNA, and from germline mutation by its higher prevalence in the population (Karki et al., 2015). Therefore, frozen tumour tissue specimens and formalin-fixed paraffin-embedded tumour tissue

specimens are frequently used as well. Focus on the allelic imbalance phenomenon also led to the justification of analysing mRNA and cDNA samples. In this case, discordance might be observed between tumour and normal tissue even if no difference was observed in the genomic primary sequence.

The choice of genotyping methods selected for association studies is dependent on the number of SNPs under investigation and the number of individuals tested. Two PCR based methods were chosen in this project: the TaqMan® SNP genotyping assay and the LightSNiP assay. The TaqMan® SNP genotyping assay is the most commonly used genotyping method in candidate gene based studies. It is designed based on 5' Nuclease assay process (see Figure 29 for the schematic diagram of 5' Nuclease assay process). For a particular biallelic SNP, probes are made for each allele with different reporter dye attached to the 5′ end and a non-fluorescent quencher at the 3′ end. In each reaction, AmpliTaq Gold DNA polymerase cleaves only the probes hybridizing to the template strands and releases the reporter dye from the suppression of the quencher molecule. The released reporter dye then becomes fluorescent and the cumulative fluorescent signals are detected by the PCR Platform (AppliedBiosystems, 2010). A substantial increase in fluorescence intensity of a reporter dye during PCR amplification indicates the existence of the corresponding allele in the test sample (see Figure 4 in **Section 2.3.2.2** for the representative raw data generated by the PCR platform).

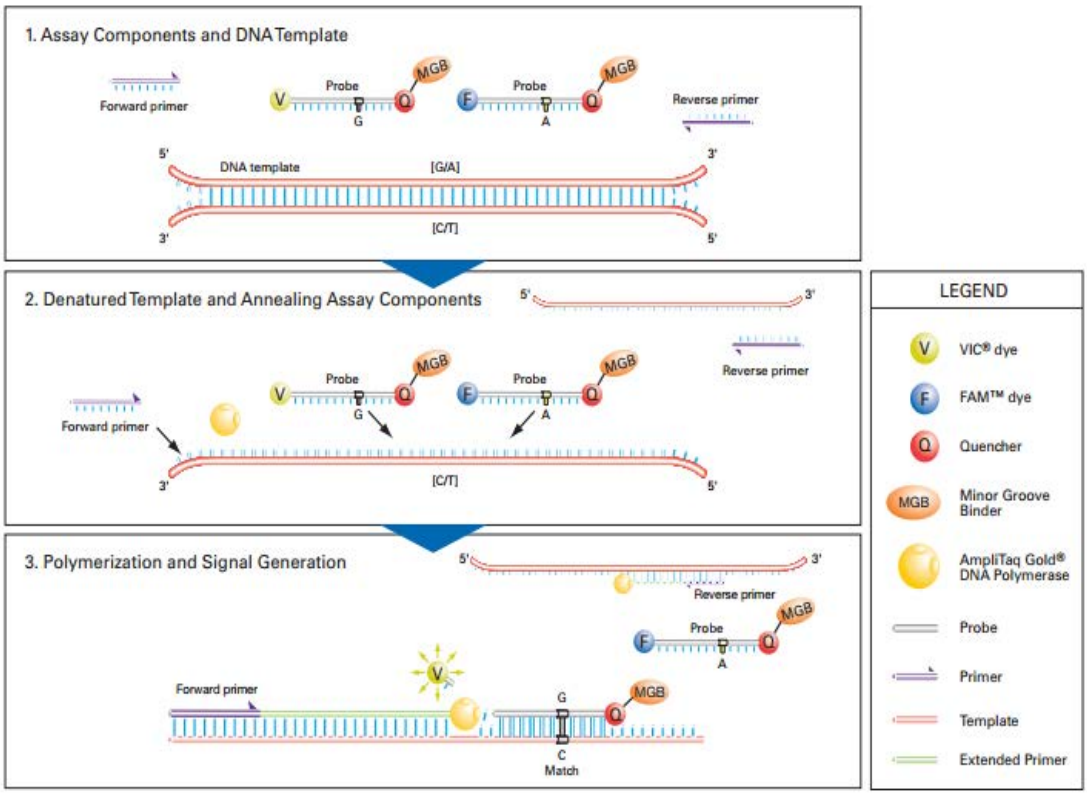


Figure 29 Schematic diagram representing the principle of the TaqMan® SNP genotyping assay

The diagram was retrieved from TaqMan® SNP Genotyping Assays protocol page 43. (http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_042998.pdf)

The LightSNiP assay had been adopted by our group as a "gold-standard" method. It was designed by TIB Molbiol (Berlin, Germany) on the basis of the SimpleProbe® and LightCycler Technology (see Figure 30 for the schematic diagram of SimpleProbe® detection). Although this genotyping method is also PCR based, the SimpleProbe® probes do not participate in the PCR amplification. A specific probe is designed to span the locus of interest with a reporter dye and a non-fluorescent quencher linked at the same end of the oligonucleotide. While hybridizing to the template DNA, the quenching effects of the probes are reduced, allowing the signals to be detected. A subsequent melting curve analysis follows the final cycle of the

PCR reactions where the PCR products are gradually heated up to 75℃ . As the temperature increases, the hybridization between the probe and the template become disturbed and the quenching effects gradually enlarge. By plotting the negative derivative of the sample fluorescence against temperature, melting peaks of each sample are generated (see Figure 5 and 6 in **Section 2.3.3** for the representative melting peaks generated by the LightSNiP assay). Perfectly matched probe and DNA template will always have higher melting peaks than mismatched probe and DNA template due to higher hybridization stability and this difference is used to determine which alleles are present in the test samples. For a biallelic SNP, homozygotes will have either the higher (if the allele present is the same as the probe oligonucleotide) or lower melting peaks (if the allele present differs from the probe oligonucleotide) while heterozygotes will have both peaks (Cheli et al., 2015).

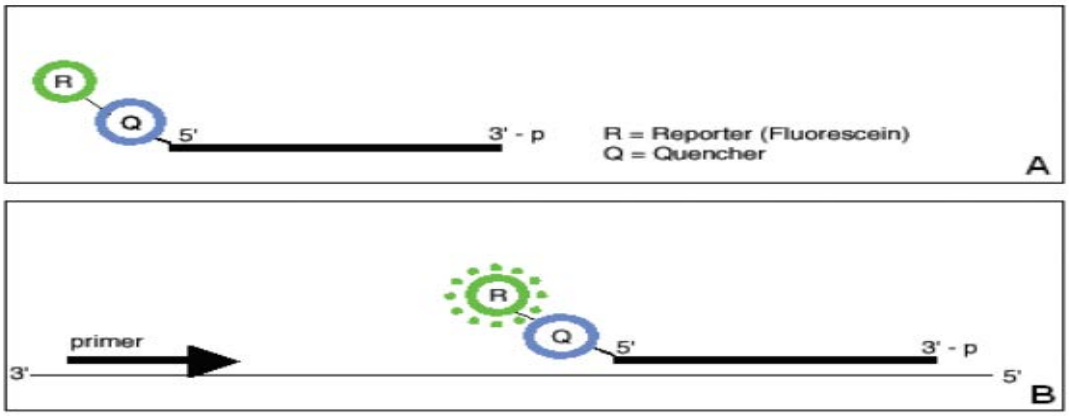


Figure 30 Schematic diagram representing the principle of the LightSNiP assay

The diagram was retrieved from Roche Applied Science Technical Note No. LC 18/2004 page5 https://lifescience.roche.com/documents/Assay-Formats-for-Use-in-Real-Time-PCR.pdf

In this project, TaqMan® SNP genotyping assay was used in both the SNP array and the subsequent validation step. In the SNP array, TaqMan® SNP genotyping assays for selected SNPs were pre-loaded and dried down in the through holes of the microscope slide–sized plate platform while for the validation of the significant findings, the single tube TaqMan® SNP genotyping assay was applied. In contrast, LightSNiP assay was only used in the validation experiments.

Before assessing the association between a particular SNP and the disease condition based on the genotyping results, a critical routine test of Hardy-Weinberg equilibrium (HWE) was performed for each SNP in the disease free controls. The Hardy-Weinberg principle was named after its discoverers Wilhelm Weinberg and G. H. Hardy. They demonstrated that when a random-mating general population was not disturbed by evolutionary forces, such as mutation, selection, genetic drift and gene migration etc., HWE would be reached and the allele and genotype frequencies of individual loci would be constant among generations (Wigginton et al., 2005). If the source population do not violate HWE, representative controls sampled from this population should also compatible with HWE while the purposively selected cases do not necessarily maintain HWE. A significant departure from HWE for a single locus observed in sampled controls often indicates a failure to reflect representativeness and homogeneity and this is may be driven by a selected population, non-random genotyping errors and non-random missing genotype data (Namipashaki et al., 2015). A goodness-of-fit chi-square test was utilized to detect the deviation between actual genotype counts and the expected HWE genotype counts, which were calculated on the basis of the frequency of each allele following

HWE law. Normally, SNPs that significantly deviated from HWE (with a p value < 0.01) were excluded from the association analyses (Lunetta, 2008).

SNPs that passed the HWE criterion were further evaluated in the association analyses. Since the underlying genetic model is unknown beforehand, statistical approaches have been proposed separately for each possible genetic model. In genetic association studies, the five most commonly seen models are co-dominant, dominant, recessive, over-dominant and log-additive inheritance models (Horita and Kaneko, 2015; Solé et al., 2006). These models are defined according to the number of copies of the risk allele needed to alter the disease risk. Binary logistic regression can be applied for all of these models in conjunction with adjustment of possible confounding variables. The co-dominant model assumes that each allele confers a different and non-additive risk. Under this model, each genotype is treated as a nominal categorical variable in a logistic regression. The dominant model assumes the risk allele confers the same risk irrespective of the numbers of risk allele a person has. In contrast, the recessive model assumes that only the homozygous carrier of the risk allele has an increased risk. In logistic regression analysis, homozygotes and heterozygotes of the risk allele are treated as the same variables for a dominant model while according to the recessive model, homozygotes of the non-risk allele and heterozygotes are treated in the same way. The over-dominant model, where homozygotes of both alleles are treated equally, is less common in the natural world, which suggests that heterozygotes have a more pronounced or better adapted phenotype than either homozygote. Unlike the models described, the log-additive model (also called multiplicative or allelic model) assesses the association based on allele counts rather than genotype frequencies and it is valid only when cases and

controls are both in HWE. It assumes that if heterozygous carriers of a risk allele confer r-fold increased risk, homozygotes of the risk allele will bear r^2-fold increased risk. This method is often criticized by geneticists as an unrealistic model since its unit of analysis is the chromosome rather than individual person, which hinders the biological interpretation of the association signals. However, it is commonly used as a primary analysis tool for association studies since the adoption of the chromosomes as test units intentionally doubles the sample size and therefore improves the chances to detect deep-hidden associations (Dorak, 2016).

Under each genetic model, the strength of the association between a particular SNP and disease outcome is measured by the odds ratio (OR). Here, a biallelic SNP with major allele M and minor allele m is used as an example. Under log-additive model, the allelic OR of m allele was estimated by the ratio of the odds of developing the disease with m allele to the odds of developing the disease with M allele[21]. Comparatively, according to the four genotype based models, genotypic OR is used rather than allelic OR. A specific genotype or the combination of two genotypes is chosen as the reference genotype depending on the model employed. Genotypic OR of a particular genotype is calculated according to the ratio of the odds of developing the disease with the genotype of interest to the odds of developing the disease with the reference genotype[22]. 95% confidence interval (CI) was calculated using the following equation: $\ln(\mathrm{OR}) \pm 1.96 \times \text{standard error of } \ln(\mathrm{OR})$ for each OR estimated, which means that if we use same sampling methods to select different sets

21 $\mathrm{OR}_{\text{m allele}} = \frac{\text{m counts in cases}}{\text{m counts in controls}} \Big/ \frac{\text{M counts in cases}}{\text{M counts in controls}}$

22 $\mathrm{OR}_{\text{genotype of interest(GOI)}} = \frac{\text{GOI Counts in cases}}{\text{GOI counts in controls}} \Big/ \frac{\text{Reference genotype counts in cases}}{\text{Reference genotype counts in controls}}$

of samples and calculate CIs, approximately 95% of the CIs calculated will include the true mean. If the 95% CI does not span the null value 1, the association is statistically significant at the 5% level. The lower limit of the 95% CI above 1 usually indicates a positive association with the disease outcome while the upper limit below one denotes a negative correlation (Dorak, 2016).

When testing the association between a single SNP and a disease outcome, a conventional 0.05 significance level cut-off is generally used. It indicates that the null hypothesis has only 5% of chance to be incorrectly rejected. However, when multiple SNPs are involved in the association study and the cut-off is still 0.05 for each test, the probability of incorrectly rejecting at least one null hypothesis can escalate considerably (e.g. Inclusion of 32 independent SNPs results in 80.6% of chance of getting at least one false positive results)[23]. To counteract the above problem, many correction methods have been suggested by researchers, such as Bonferroni (Dunnett, 1955), Holm (Holm, 1979), Hommel (Hommel, 1988), Hochberg (Hochberg, 1988), Benjamini & Hochberg (Benjamini and Hochberg, 1995) and Benjamini and Yekutieli (Benjamini and Yekutieli, 2001).

The two most commonly used methods are the Bonferroni method and the Benjamini & Hochberg method, respectively. The Bonferroni method is the most stringent method for multiple comparisons, which corrects the p value by multiplying it by the total number of tests. After Bonferroni correction, if a significance cut-off of 0.05 is used, it expects that the probability of incorrectly rejecting at least one null hypothesis is 5%. Although this method minimizes false-positive associations

23 Calculated by $1-(1-0.05)^{32}$

identified, it also minimizes the power to detect genuine associations. Benjamini & Hochberg method is also called false discovery rate (FDR) method. Compared with Bonferroni method, FDR method is preferred by geneticists since it strikes a good balance between false positive and false negative errors. This method can be performed by programming software R using command **p.adjust (p, method="fdr", length (p))**[24], which was written by Prof Gordon Smyth[25] based on the re-interpretation of both Benjamini & Hochberg and Benjamini and Yekutieli (BY) procedures. It ranks the raw p values from smallest to the largest and then corrects each p value using formula: adjusted p value=raw p value $\times \frac{\text{total number of tests}}{\text{the rank of the raw p value}}$. Sometimes the adjusted p value of a target with a smaller raw p value is larger than that for its less significant counterparts. In order to maintain the monotonicity, a second step was used to replace the non-monotonic adjusted p values with the smallest adjusted p value of all the less significant targets in original analysis. Conventionally, a false discovery rate cut-off of 0.15 is used, which indicates 15% of the statistically significant associations identified are really false positives.

In this project, a log-additive model was chosen for initial screening of the genotyping results and crude (unadjusted) allelic OR was obtained for each SNP using Pearson's $\chi 2$ test (if all expected counts are ≥ 5) or Fisher's exact test (If any expected counts are <5). After multiple testing corrections with the FDR method (with FDR cut-off of 0.15), significant SNPs were further validated and examined under all five genetic models using age-adjusted binary logistic regression. In order to determine how well each model fitted the genotyping data, a likelihood ratio test

24 https://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html
25 Prof Gordon Smyth, Division Head at Walter and Eliza Hall Institute of Medical Research, The University of Melbourne

(LRT) was performed for each model in comparison with the null model (only including the intercept)(González et al., 2007). When the LRT test was not sensitive enough to determine the best-fit model, both Akaike's information criterion (AIC) and Bayesian information criterion (BIC) scores were calculated so as to find the model that minimized the information loss (Akaike, 1974; Burnham and Anderson, 2003). The model with lowest AIC or BIC scores were preferred (González et al., 2007; Solé et al., 2006).

As with other experimental designs, genetic association studies require an adequate sample size to capture the genuine associations. In fact, a sample size estimation or power calculation should be considered a priority, which only surpassed by the establishment of the research question and the selection of the proper sample type. In this project, Quanto developed by Jim Gauderman and John Morrison was selected for sample size estimation under the assumption of the log-additive model (Gauderman, 2006). Please see **Section 2.3.2.1** and **Appendix 1** for the estimation of the sample size under the log-additive model. When multiple SNPs are involved in the study, the sample size required is larger than that calculated for a single SNP, which can be calculated by multiplying the estimated sample size by a factor: $\log_{10}(\text{number of SNPs to be genotyped})$ (Stram, 2014). In this study, the factor used was roughly 1.5[26]. Therefore an achievable sample size of 475 was selected for controls and cases for this project, which reconciled with the format of the PCR platform and gave enough power under the majority of conditions for log-additive model.

26 Factor=$\log_{10} 32$

In order to precisely determine the genotypes of each sample in an array based study, a validation step is usually performed to confirm significant findings by genotyping the same set of samples with a different genotyping method. This kind of validation mainly focuses on the analytical validity of the genotyping results. Another kind of validation which can be performed is to assess another set of samples from the original population or a set of samples from a different population to confirm the generalizability of the associations identified. The latter usually requires an even larger sample collection to overcome the possible overestimation of effect size in the initial study (a phenomenon known as winner's curse). Within the timeframe of the project, only the former could be conducted. The validated association could then be interpreted in one of three possible ways, including direct association (i.e. caused directly by the typed causal SNP), indirect association (i.e. caused by an unknown causal SNP in high LD with the typed SNP itself) or spurious association (i.e. caused by chance or artefacts). Although a lot of prediction tools have been developed to further evaluate significant findings, it is challenging to tease out direct associations from the other two situations unless further validation (or functional evidence) is provided. In this project, rational selection of the candidate SNPs according to their functional potential to some extent increased the possibility to detect direct associations.

In complex disease with a long natural history and progression process, the role of a SNP might not be the same during the different disease stages. By genotyping a particular SNP in samples belonging to different disease categories or stages, a better understating of the genetic associations between this SNP and a broad spectrum of transition states could be obtained. Good examples have been given previously

(Safaeian et al., 2012; Wang et al., 2010). These studies investigated the associations between candidate SNPs and two transition states of cervical cancer (i.e. viral persistence and progression to pre-cancer/cancer) and successfully identified SNPs that specifically corresponded to each state and SNPs associated with both states. In this project, thanks to Dr Michelle Etherson and Luhan Jiang, significant SNPs identified from the SNP array were genotyped in additional samples with HPV-related diseases and the results were pooled with the array results. The inclusion of additional samples increased the robustness of the association analysis.

In general, the aim of this chapter is to assess the associations between candidate SNPs and HPV-related diseases and to further validate the significant findings using different genotyping methods. The general results of the SNP array were described in **Section 4.2**. The validations of the significant results were described in **Section 4.3** and **Section 4.4**.

4.2 Overview of the SNP array

In total, 27 out of 32 SNPs yielded analyzable results in the array. TMC8 rs16970849, ACKR2 rs4683336，NEIL2 rs804270, IL-8 rs4073 and PGR rs11224561 were excluded from analysis since they were either shown to be monomorphic or had a significant departure from Hardy-Weinberg equilibrium in the controls. DNA extracted from FFPE samples performed poorly in the array compared to that extracted from LBC samples, especially those extracted from FFPE cervical cancer samples. Therefore, genotyping results of all the cancer samples (n=62) were omitted in the initial analysis. The actual cases used in the initial analysis only consisted of HR-HPV positive high-grade squamous intraepithelial lesions. Characteristics of the

case group (n=413) and control group (n=475) in the initial analysis of the SNP array are provided in Table 19. As shown in this table, the mean age of the cases was significantly younger than that of controls, which indicated that age adjustment was necessary in the final association analysis. Sample sources of the cases also significantly differed from the controls, although this was were very unlikely to affect the genotyping results and the subsequent analysis.

Table 19 Characteristics of the case and control group in the initial analysis

Variables	HR-HPV positive HSIL Cases (n=413)	Control(n=475)	p values
Age, y (mean ± SD)	20.12 ±0.51	38.18±11.43	<0.0001[a]
Source, n (%)			<0.0001[b]
FFPE	149(44.4%)	0(0%)	
LBC samples	264(55.6%)	475(100%)	

a. Two-sided Mann Whitney test for difference between cases and controls;

b. Two-sided Fisher's exact test for distributions between cases and controls.

Genotyping results for all the samples tested are shown in Table 20. The highest and lowest call rates obtained for the 27 analyzable SNPs were 99.32% and 76.01% respectively. The frequency of the minor allele homozygotes was found to be lower than 1% in our sample set for three SNPs: rs17849071, rs1799724 and rs361525.

Table 20 OpenArray genotyping report for 27 analyzable SNPs

rs12452890	98.31%	yes	10	116	693	69
rs17849071	86.60%	minor allele homozygotes<0.01	3	113	653	119
NCBI reference	call rate	Have all three genotypes	AA	AB	BB	undetermined
NCBI reference	call rate	Have all three genotypes	AA	AB	BB	undetermined
rs1799724	96.30%	minor allele homozygotes<0.01	746	403	607	24
rs1042838	97.30%	yes	14	243	607	24
rs1800629	98.99%	yes	736	279	565	39
rs1807137	80.82%	yes	512	229	384	117
rs2506542	76.01%	yes	344	278	58	213
rs2069705	98.98%	yes	364	360	196	62
rs2069718	79.82%	yes	263	345	420	180
rs2104286	95.50%	yes	465	335	448	150
rs2227956	84.89%	yes	478	249	664	135
rs2600887	95.72%	yes	479	310	61	38
rs2736340	97.32%	yes	694	215	16	75
rs2830656	95.47%	yes	468	257	425	129
rs3212986	97.36%	minor allele homozygotes<0.01	68	215	729	37
rs3217810	98.40%	yes	69	351	426	50
rs3760896	99.20%	yes	89	330	454	175
rs4149742	99.32%	yes	694	252	456	76
rs4377742	99.09%	minor allele homozygotes<0.01	684	410	286	36
rs4646296	97.25%	yes	89	238	498	464
rs4713	91.15%	yes	87	238	420	121
rs4880	99.69%	yes	184	430	238	156
rs4971	98.96%	yes	97	264	370	98
rs5023	95.65%	yes	182	216	226	464
rs7072793	96.15%	yes	156	420	278	34
rs8042703	96.30%	yes	146	300	228	299
rs8043703	96.37%	yes	156	420	278	284
rs853360	97.64%	yes	462	341	64	21

To assess the associations between analyzable SNPs and HPV-related high-grade squamous intraepithelial lesion (HSIL), raw p value and FDR adjusted p value were calculated for each SNP under the log-additive model as described in **Section 2.3.2.2.** Please see table 21 for the list of raw p values and FDR adjusted p-values under the log-additive model.

Table 21 Genotyping results for 27 analyzable SNPs in crude analysis (a list of raw p values and FDR adjusted p-values under the log-additive model)

NCBI reference	Raw.p (log-additive model)	FDR
rs2234671	0.0001	0.0027
rs2623047	0.0075	0.10125
rs17849071	0.014	0.126
rs7072793	0.028	0.189
rs1126579	0.099	0.5346
rs2790	0.26	0.838421
rs2104286	0.27	0.838421
rs34137742	0.3	0.838421
rs1042838	0.31	0.838421
rs361525	0.32	0.838421
rs12452890	0.45	0.838421
rs2596542	0.47	0.838421
rs2107538	0.5	0.838421
rs266085	0.52	0.838421
rs853360	0.52	0.838421
rs3760396	0.54	0.838421
rs1800629	0.56	0.838421
rs1799724	0.57	0.838421
rs10774671	0.59	0.838421
rs4074	0.66	0.891
rs11556218	0.71	0.912857
rs2857656	0.78	0.93913
rs1801157	0.8	0.93913
rs2069705	0.88	0.9504
rs2699887	0.88	0.9504
rs3921	0.94	0.96
rs3117604	0.96	0.96

In the crude analysis, allele frequencies of 4 SNPs, including PIK3CA rs17849071, CXCR1 rs2234671, SULF1 rs2623047 an IL2RA rs2072793, exhibited significant differences between HR-HPV positive HSIL cases vs. cytologically normal, HR-HPV negative controls (with a raw p value< 0.05). The four SNPs intentionally included due to their associations with cervical cancer were not found to be associated with HR-HPV positive HSIL[27]. The associations of the CXCR1 rs2234671, SULF1 rs2623047 and PIK3CA rs17849071 remained after multiplicity adjustment (FDR<0.15). Accordingly, SULF1 rs2623047 and CXCR1 rs2235671,

27 TMC8 rs16970849 was monomorphic in our sample set while the other three was found to be insignificant in the initial analysis.

the two most significant SNPs identified in crude analysis were chosen as targets for further validation and functional analysis. PIK3CA rs17849071 was not followed in this project due to its low frequency of the minor allele homozygotes in the sample set.

4.3 The association between SULF1 rs2623047 and HR-HPV-related diseases

4.3.1 Analytical validity of the array results for SULF1 rs2623047

For SULF1 rs2623047, 707 samples had valid genotype calls in the original array. 656 of these 707 samples were picked and re-genotyped using LightSNiP assay and/or single tube TaqMan® SNP Genotyping assay[28] as described in **Section 2.3.3** and **Section 4.1**) according to their availability. In total, 139 samples were verified by all three genotyping methods (i.e. SNP array, LightSNiP assay and single tube TaqMan® SNP Genotyping assay). Genotyping results were relatively consistent among the three methods and only 5 samples (out of 656) had discordant results (i.e. inconsistent between any two methods). Although the call rate of this SNP (80%) was just above the threshold for the array, the overall concordance was satisfactory (99.2%). In order to check whether the significance was caused by a high no-call rate, all the no-call samples (888-707=181) in the array were re-genotyped using LightSNiP assay.

Genotyping results for 475 controls and HR-HPV positive HSIL cases are summarized in Table 22. Both controls and cases were in HWE, with HWE $\chi 2$ p-value of 0.64 and 0.47 respectively. Significant associations between this SNP and HR-HPV positive HSIL were observed under four genetic models after age

28 Also performed by WTCRF Genetics core

adjustment, including over-dominant model, dominant model, recessive model and log-additive model (see Table 23 for details). According to the AIC and BIC values, the best-fit model for this SNP was the Log-additive model (OR: 0.50, 95%CI 0.32-0.77), suggesting a possible multiplicative protective effect of G alleles against HR-HPV-related HSIL.

Table 22 SULF1 rs2623047 genotype frequencies for all the subjects and separately for controls and HR-HPV positive HSIL cases.

SULF1 rs2623047 genotype frequencies (Total n=888)				
Genotype, n (%)	**All subjects**	**Controls**	**HR-HPV positive HSIL Cases**	
A/A	**314(36%)**	**155(33%)**	**159(38.8%)**	
A/G	**422(48%)**	**235(50%)**	**187(45.6%)**	
G/G	**146(17%)**	**82(17%)**	**64(15.6%)**	
No call	**6**	**3**	**3**	**---**

Table 23 The association between SULF1 rs2623047 and HR-HPV positive HSIL under five different genetic models (adjusted by age)

Model	Genotypes	Controls	HR-HPV positive HSIL Cases	OR (95%CI)	p-value	AIC	BIC
	A/A	153(32.5%)	159(38.8%)	1			
Co-dominant	A/G	235(50%)	187(45.6%)	0.45(0.21-0.96)	0.0058	305.2	324.3
	G/G	82(17.4%)	64(15.6%)	0.25(0.10-0.60)			
Dominant	A/A	154(32.7%)	159(38.8%)	1	0.0052	305.7	320.1
	A/G+G/G	317(67.3%)	251(61.2%)	0.38(0.18-0.78)			
Recessive	A/A+A/G	389(32.6%)	346(84.4%)	1	0.017	307.8	322.1
	G/G	82(17.4%)	64(15.6%)	0.42(0.21-0.83)			
Over-dominant	A/A+G/G	236(50.1%)	223(54.4%)	1	0.51	313.1	327.4
	A/G	235(49.9%)	187(45.6%)	0.82(0.45-1.48)			
Log-additive	---	---	---	0.50(0.32-0.77)	0.0014	303.1	317.7

This table depicts the association between SULF1 rs2623047 genotypes and HR-HPV positive HSIL. Data displayed as n (%cases). OR, 95%CI, p-value, AIC and BIC were calculated for each inheritance model. Samples with missing genotype data or age records were automatically removed from the statistical analysis by the SNPstats software.

4.3.2 SULF1 rs2623047 is associated with HR-HPV infection rather than the development of high-grade squamous intraepithelial lesion and cancer

Unlike the SNP array and the single tube TaqMan® SNP Genotyping assay, the LightSNiP assay required less genomic DNA to make an accurate genotype call and showed good compatibility with FFPE samples. Therefore, this method was applied to re-genotype the omitted FFPE cancer samples in the array sample set. Due to limited sample volume, valid genotype results for this SNP were finally obtained for 39 of the 62 cancer samples. As described in **Section 2.3.3**, genotyping results of an additional 83 HR-HPV positive cytologically normal LBC samples and 21 HR-HPV positive cervical cancer samples annotated as a consequence of work performed by Luhan Jiang and Dr Michelle Etherson were pooled with the array results. Characteristics and genotyping results of the control group and different disease categories are summarized in Table 24. HR-HPV positive cytologically normal group and HR-HPV positive cervical cancer group were also in HWE, with HWE $\chi 2$ p-value of 0.78 and 0.81 respectively.

Table 24 Characteristics and genotyping results of the control group and different disease categories for SULF1 rs2234671

Variables	Control (n=475)	HR-HPV+/Cyt- (n=83)	HR-HPV+ HSIL (n=413)	HR-HPV+ cancer (n=83)
Age, y (mean ± SD)	38.18±11.43	20±0	20.12 ±0.51	48.27±16.07
Genotype, n (%)				
A/A	155(33%)	37(48.1%)	159(38.8%)	23(38.3%)
A/G	235(50%)	32(41.6%)	187(45.6%)	29(48.3%)
G/G	82(17%)	8(10.4%)	64(15.6%)	8(13.3%)
No call	3	6	3	23

Similar to HR-HPV positive HSIL, the HR-HPV positive, cytologically normal group was also significantly different from the HR-HPV negative, cytologically normal group under four genetic models after age adjustment (see Table 25 for details).

Table 25 Association of SULF1 rs2623047 in HR-HPV positive cytologically normal cases versus HR-HPV negative cytologically normal controls under five different genetic models (adjusted by age)

Model	Genotypes	Controls	HR-HPV+/Cyt- cases	OR (95%CI)	p-value	AIC	BIC
	A/A	153(32.5%)	37(48%)	1			
Co-dominant	A/G	235(50%)	32(41.6%)	0.30(0.12-0.76)	0.0009	155.6	172.8
	G/G	82(17.4%)	8(10.4%)	0.12(0.04-0.42)			
Dominant	A/A	154(32.7%)	37(48%)	1	0.0007	156.2	169.1
	A/G+G/G	317(67.3%)	40(52%)	0.24(0.10-0.58)			
Recessive	A/A+A/G	389(32.6%)	69(89.6%)	1	0.0082	160.6	173.5
	G/G	82(17.4%)	8(10.4%)	0.26(0.09-0.73)			
Over-dominant	A/A+G/G	236(50.1%)	45(58.4%)	1	0.24	166.2	179.1
	A/G	235(49.9%)	32(41.6%)	0.63(0.30-1.35)			
Log-additive	---	---	---	0.34(0.19-0.63)	0.0002	153.7	166.7

This table depicts the association between SULF1 rs2623047 genotypes and HR-HPV infection. Data displayed as n (%cases). OR, 95%CI, p-value, AIC and BIC were calculated for each inheritance model. Samples with missing genotype data or age records were automatically removed from the statistical analysis by the SNPstats software.

The G allele of this SNP was shown to be associated with decreased susceptibility to HR-HPV infection under the best fit model: log-additive model (OR: 0.34 95%CI 0.19-0.63). However, there was no significant difference observed in the allele or genotype frequencies of SULF1 rs2623047 among HR-HPV positive cytologically normal group, HR-HPV positive HSIL group and HR-HPV cervical cancer group, which suggested that this SNP was not related the development of high-grade squamous intraepithelial lesion and the progression to cervical cancer.

4.4 The association between CXCR1 rs2234671 and HR-HPV-related diseases

4.4.1 Analytical validity of the array results for CXCR1 rs2234671

For CXCR1 rs2234671, 811 samples had valid genotype calls in the array. All except two were re-genotyped using either LightSNiP assay or single tube TaqMan® SNP genotyping assay. In total, 572 samples were genotyped with all three methods. The concordance between LightSNiP assay and single tube TaqMan® SNP Genotyping assay was 99.0%. However, compared with these two methods, 56 samples (out of 572) had different calls in the SNP array. Genotyping data of this SNP was thereby sent to the Technical support team at ThermoFisher Scientific and a bridging issue was identified subsequently. The details were not clear but I was advised to discard all the results for this SNP in the original array due to low analytical validity.

4.4.2 The minor allele frequency of CXCR1 rs2234671 was significantly increased in HR-HPV positive cancer cases, especially in HPV16 positive cancers

Since both LightSNiP assay and single tube TaqMan® SNP Genotyping assay were shown to give reproducible results for this SNP, the whole set of samples in the array for this SNP were re-genotyped using either of these two methods. Meanwhile, an additional 83 HR-HPV positive cytologically normal LBC samples, 21 HR-HPV positive cervical cancer samples, 129 HR-HPV positive vulval intraepithelial neoplasia (VIN) samples and 23 HR-HPV positive vulval cancer samples were genotyped for this SNP by Luhan Jiang and Dr Michelle Etherson with results pooled in the association analysis. Characteristics and genotyping results of the control group and different disease categories are summarized in Table 26.

Table 26 Characteristics and genotyping results of the control group and different disease categories for CXCR1 rs2234671

Variables	HR-HPV-/Cyt- cervical samples (n=475)	HR-HPV+/Cyt- cervical samples (n=83)	HR-HPV+ cervical HSIL (n=413)	HR-HPV+ cervical cancer (n=83)	HR-HPV+ VIN (n=129)	HR-HPV+ vulval cancer (n=23)
Age, y (mean ± SD)	38.18±11.43	20±0	20.12 ±0.51	48.27±16.07	47.5±12.4	56.6±12.2
Genotype, n (%)						
G/G	410(87%)	70(85%)	360(89%)	64(79%)	103(83%)	15(65%)
G/C	58(12%)	12(15%)	45(11%)	16(20%)	20(16%)	7(30%)
C/C	3(1%)	0(0%)	1(0%)	1(1%)	1(1%)	1(4%)
No call	4	1	7	2	5	0
Allele, n (%)						
G	878(93.2%)	152(92.7%)	765(94.2%)	144(88.9%)	226(91.1%)	37(80.4%)
C	64(6.8%)	12(7.3%)	47(5.8%)	18(11.1%)	22(8.9%)	9(19.6%)

All the groups were in HWE, with HWE $\chi 2$ p-value of 0.55, 0.47, 0.74, 1, 0.98 and 0.87 respectively. No significant difference was found in the allele frequencies of CXCR1 rs2234671 among HR-HPV negative, cytologically normal control group (G 93.2%, C 6.8%), HR-HPV positive cytologically normal group (G 92.7%, C 7.3%), HR-HPV positive cervical HSIL group (G 94.2%, C 5.8%) and HR-HPV positive VIN group (G 91.1%, C 8.9%). Also, the genotype counts did not vary among these groups. However, significantly enriched C allele was observed in HR-HPV positive cervical cancer cases and vulval cancer cases, compared with HR-HPV positive cervical HSIL cases and HR-HPV positive VIN respectively. Pearson's $\chi 2$ test p value was 0.0132 for the former and 0.0301 for the latter.

Since the small sample size in some of the individual groups was not well-fitted for analysis based on genotypes, the two cancer groups (cervical and vulval) were merged and the results of the HR-HPV positive cervical HSIL group were combined

with the HR-HPV positive VIN group. An interesting finding was obtained by analysing the data set this way. When stratifying each disease category using HPV status, the enrichment of C allele was only observed in HPV16 positive cancers (see Figure 31).

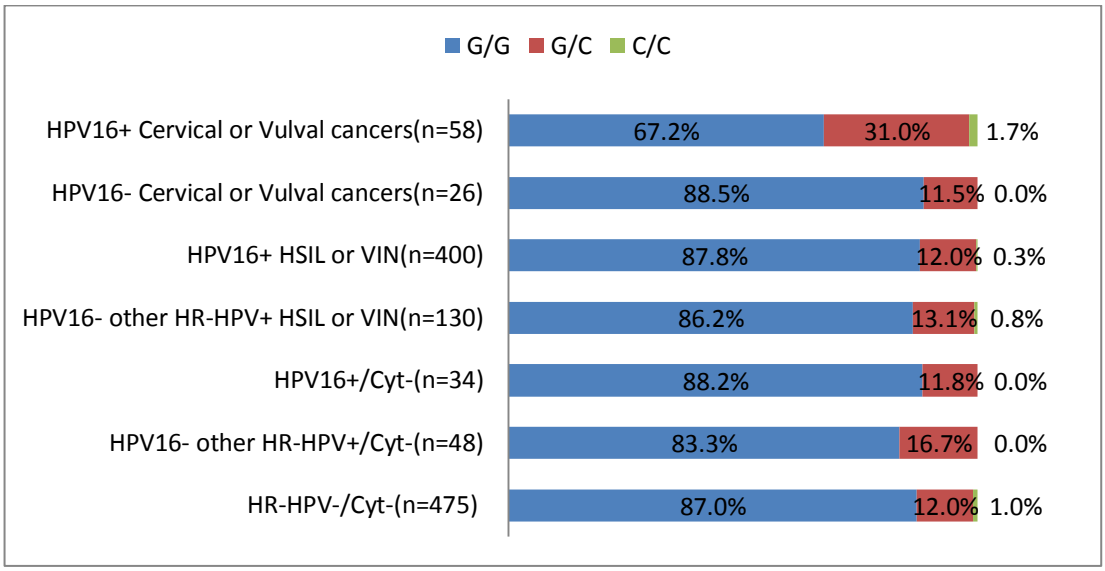


Figure 31 Genotype frequencies of CXCR1 rs2234671 in the HR-HPV negative, cytologically normal control group and other disease categories (stratified by HPV16)

As can be seen from the above figure, the frequency of the C allele carriers was significantly higher in the combined HPV16 positive cancer group, compared with the HPV16 positive HSIL or VIN group (32.7% vs. 12.3%, Pearson's $\chi 2$ test $p<0.0001$) or the HR-HPV negative, cytologically normal control group (32.7% vs. 13%, Pearson's $\chi 2$ test $p<0.0001$). Apart from the combined HPV16 positive cancer group, no significant difference was found in genotype frequencies or allele frequencies of CXCR1 rs2234671 among all other groups.

The association between this SNP and the progression of HPV16 related cancer was evaluated by comparing the combined HPV16 positive cancer group with the HPV16

positive HSIL or VIN group controls under different genetic models after age adjustment (see Table 27 for details). The CXCR1 rs2234671 SNP was shown to be associated with the development of HPV16 related cancer under 4 of the 5 genetic models and the best-fit model for this SNP was the log-additive model (OR: 3.23, 95%CI 1.56-6.67). The likelihood ratio test p value for this model is 0.0018.

Table 27 The association between CXCR1 rs2234671 and the progression of HPV16 positive cancers under five different genetic models (adjusted by age)

Model	Genotypes	HPV16+ cervical HSIL and VIN cases	HPV16+ cancers	OR (95%CI)	p-value	AIC	BIC
	G/G	351(87.8%)	38(66.7%)	1			
Co-dominant	C/G	48(12%)	18(31.6%)	3.22(1.48-7.02)	0.0078	242.6	259.1
	C/C	1(0.2%)	1(1.8%)	10.7(0.35-329.40)			
Dominant	G/G	351(87.8%)	38(66.7%)	1	0.0024	241	253.4
	C/G+C/C	49(12.2%)	19(33.3%)	3.37(1.56-7.25)			
Recessive	G/G+C/G	399(99.8%)	56(98.2%)	1	0.23	248.8	261.2
	C/C	1(0.2%)	1(1.8%)	8.34(0.27-254.12)			
Over-dominant	G/G+C/C	352(88%)	39(68.4%)	1	0.0049	242.4	254.7
	C/G	48(12%)	18(31.6%)	3.13(1.44-6.81)			
Log-additive	—	—	—	3.23(1.56-6.67)	0.0018	240.6	252.9

This table depicts the association between CXCR1 rs2234671 genotypes and the progression of HPV 16 positive cancers, compared with the HPV16+ HSIL or VIN group. Data displayed as n (%cases). OR, 95%CI, p-value, AIC and BIC were calculated for each inheritance model. Samples with missing genotype data or age records were automatically removed from the statistical analysis by the SNPstats software.

4.5 Discussion

The individual susceptibility to the development of cervical cancer is considered to be associated with genetic variations in specific genes. In this chapter, 32 rationally selected SNPs within 22 genes were genotyped to determine their association with HR-HPV-related diseases and 27 of them yielded analyzable genotyping results.

Unlike the majority of candidate gene based association studies and GWAs, genomic DNA extracted from LBC and FFPE samples, instead of routinely-used blood samples, was used in this project for SNP genotyping. As mentioned in **Section 4.1**, theoretically, this should not cause any changes to the genotyping results. Vieira et al previously showed that the Taqman® OpenArray® genotyping assay could be used to genotype genomic DNA extracted from FFPE samples (Vieira et al., 2013). All the FFPE cervical cancer samples in my array sample set were freshly extracted using QIAGEN GeneRead DNA FFPE Kit, which could enzymatically remove artificial C>T mutations caused by cytosine deamination to maximise yield and purity of genomic DNA from FFPE samples suitable for next-generation sequencing applications. In our group, quality control of the extracted DNA was performed by measuring both 260/230 nm and 260/280 nm absorbance ratios using a NanoDropTM 1000 spectrophotometer. Meanwhile, quality control of the extracted DNA was also performed by a PicoGreen dsDNA quantitation assay (ThermoFisher Scientific) in WTCRF Genetic core. However, although these samples passed the quality control by both laboratories, few of them had valid genotyping results in the SNP array and they also fared poorly in the single tube TaqMan® SNP Genotyping assay in the validation step (data not shown). These findings suggested that genomic DNA extracted from FFPE samples may not be suitable for the Taqman® SNP genotyping system.

In spite of the poor performance of genomic DNA from FFPE samples in the array, genotyping results of genomic DNA from LBC samples were very promising. LBC samples have been shown to be reliable for the application of molecular techniques, such as fluorescence in situ hybridization, methylation-specific PCR and denaturing

high-performance liquid chromatography (Li et al., 2014; Spathis et al., 2011; Sun et al., 2012). However, not many genetic association studies have been performed with genomic DNA extracted from LBC samples (Casalicchio et al., 2014; Mehta et al., 2015). As cervical screening programmes become more and more widespread, LBC samples might provide an alternative to blood tests for genetic marker detection. Previous work done in our group indicated that properly-stored LBC samples could yield DNA of good quality and quantity. Genotyping results in the array further supported the utility of genomic DNA extracted from LBC samples in TaqMan® SNP genotyping system. In the validation step, the LightSNiP assay was applied, which showed many advantages over single tube TaqMan® SNP Genotyping assay. Much less genomic DNA (i.e. ~10ng/reaction) was required while a much higher call rate was reached. Most surprisingly, even problematic FFPE extracts had valid and reproducible results in LightSNiP assay. The application of the LightSNiP assay effectively avoided the potential pitfalls caused by wrong call and no call samples.

In the SNP array, I intentionally included 5 SNPs that have been shown to be associated with cervical cancer in other studies so as to check the generalizability of these associations in Scottish population. However, since all the cancer samples were excluded from the initial analysis, these associations could not be reproduced in this project. Among these five SNPs, TMC8 rs16970849 was excluded from the analysis because all the subjects in our sample set were homozygotes of the major allele. Technical validation of the other four SNPs, including rs361525, rs1800629, rs853360 and rs34137742, was not performed due to the time and cost constraints of this project. In spite of this, the call rates of these three SNPs were all acceptable (see Table 20) and the genotype groups were well separated in the allelic discrimination

plot. According to the initial analysis, the allele frequencies of these four SNPs were not significantly different between HR-HPV negative cytologically normal controls and those HR-HPV positive HSILs, which suggested that these SNPs were not associated with the development of the HR-HPV positive HSIL (see Table 20). This project is the first to assess the associations between HR-HPV-related squamous intraepithelial lesions and rs853360 or rs34137742. Although controversial observations of the association between TNF rs1800629 and cervical cancer were found in published studies, this SNP was consistently shown to be unrelated to the development of squamous intraepithelial lesions (Li et al., 2013; Rotar et al., 2014; Sousa et al., 2014). The results obtained in this project provided further evidence for the lack of association of rs1800629 with HR-HPV-related HSIL. Similarly, results of rs361525 in this project were consistent with a previous finding in a Mexican population (Nieves-Ramirez et al., 2011). However, firm conclusions could not be drawn without further verification and validation.

Intriguingly, two SNPs, SULF1 rs2623047 and CXCR1 rs2234671, were found to be significantly associated with HR-HPV related diseases and were subsequently followed-up. They were genetically linked to two different transition states during the natural history of the disease. SULF1 rs2623047 revealed a strong significant association with the susceptibility to HR-HPV infection while CXCR1 rs223671 was solely associated with the development of HR-HPV-related cancers, especially in HPV16+ cancer cases.

SULF1 rs2623047 is located upstream of the sulfatase 1 (SULF1) gene. This gene encodes a heparan sulfate modifying enzyme, which is involved in the activation of many signalling pathways by selectively modifying heparan sulfate proteoglycans

(HSPGs) (Pascale et al., 2016). The role of SULF1 in cancers is not entirely clear according to the current literature. A cell type-specific impact of SULF1 on carcinogenesis has been suggested (Bret et al., 2011; Chen et al., 2009; Lai et al., 2008; Yang et al., 2011). SULF1 has also been shown to be important in host-pathogen interactions. Kim et al showed that over-expression of SULF1 in HeLa cells inhibited Chlamydia binding through HSPG 6-O-desulfation (Kim et al., 2013). While the receptor for HPV is poorly characterized, viral entry has also been shown to be facilitated by HSPGs (Horvath et al., 2010; Kines et al., 2016; Raff et al., 2013). However, the specific roles of SULF1 in HPV persistence and the development of HR-HPV-related cervical cancer have not been investigated comprehensively. In the previously mentioned array-based study performed by Wang and colleagues, SNPs in the SULF1 gene were found to be significantly associated with HPV persistence (HPV persistence vs. Random control, P=0.0056). No significant differences were observed between the HPV persistence group and the cervical pre-cancer/cancer group (CIN3/cervical cancer vs. HPV persistence, P=0.6935). Within the gene, three intronic tagSNPs were found to reach the significance threshold individually in the array, which were rs4737999, rs4284050 and rs10108002 respectively. All these three SNPs were strongly associated with HPV persistence but did not differ between the HPV persistence group and the cervical pre-cancer/cancer group (Wang et al., 2010). Our findings for SULF1 rs2623047 added to the evidence that SULF1 played a pivotal role in the pathogenesis of HR-HPV infection. It is noteworthy that none of the tagSNPs in the published studies, including those failed to reach significant levels, was in high linkage disequilibrium with rs2623047, which indicated that SULF1 rs2623047

contributed to HR-HPV infection risk genetically independent of the previously reported susceptibility loci.

According to the Phase 3 1000 Genomes Browser based on Ensembl v80 GRCh37[29], SULF1 rs2623047 locates at chromosome 8, the promoter region of the SULF1 gene. Two alleles, A and G, exist at this locus. The G allele was shown to be the major allele in African and East Asian populations and considered to be the ancestral allele. However, in South Asian, American and European populations, the A allele became the major one (Consortium, 2015). The allelic distribution of this SNP in our HR-HPV negative cytologically normal control group is almost the same as that of the British population in the 1000 Genomes database. To date, only two studies on SULF1 rs2623047 have been published (Han et al., 2011; Okolicsanyi et al., 2014). Han et al showed that rs2623047 AA genotype was related to shorter progression-free survival in ovarian cancer. As a promoter SNP, rs2623047-G allele was found to confer a 1.4-fold higher promoter activity than A allele in SKOV3 cell line (Han et al., 2011). The study performed by Okolicsanyi also suggested rs2623047-A allele was a risk allele for breast cancer. In this study, rs2623047 alleles were reported in reverse orientation to genome. C actually means G and T corresponds to A (Okolicsanyi et al., 2014). Our genotyping results consistently indicated that carriers of rs2623047-A were at increased risk of developing HR-HPV infection, although the underlying biological mechanisms involved in cancer and viral infection are clearly very different. Apart from for the study performed by Han et al, no functional evidence published to elucidate a mechanism behind the observations. This SNP was not selected for functional follow-up within the

29 http://phase3browser.1000genomes.org/index.html

timeframe of this project. However, the functional effects of this SNP and potential functional SNPs in high LD with this SNP could be roughly predicted using online tools, such as HaploReg v4[30], F-SNP[31], SNPinfo Web Server[32] or PROMO[33]. For example, both HaploReg v4.1 and SNPinfo Web Server predicted that rs2623047 could affect transcription factor binding. Also, HaploReg v4 provided a list of SNPs in high LD with rs2623047, which were potential candidates for further functional analysis. More functional predictions of this SNP are described in **Section 6.2.2**.

Another significant SNP identified in this study was found within CXCR1 gene, which encodes a G protein–coupled receptor (GPCR). This receptor had been shown to be a high affinity receptor for IL-8 (Griffith et al., 2014). CXCR1 rs2234671 is a non-synonymous biallelic SNP located at the coding region of the gene (see Figure X for the position of rs2234671). The major allele is G and the G-C substitution leads to an amino acid substitution from serine (AGC) to threonine (ACC) at the third extracellular loop of the encoded receptor (see Figure 32 for the amino acid position affected by rs2234671).

30 http://archive.broadinstitute.org/mammals/haploreg/haploreg.php
31 http://compbio.cs.queensu.ca/F-SNP/
32 https://snpinfo.niehs.nih.gov/
33 http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3

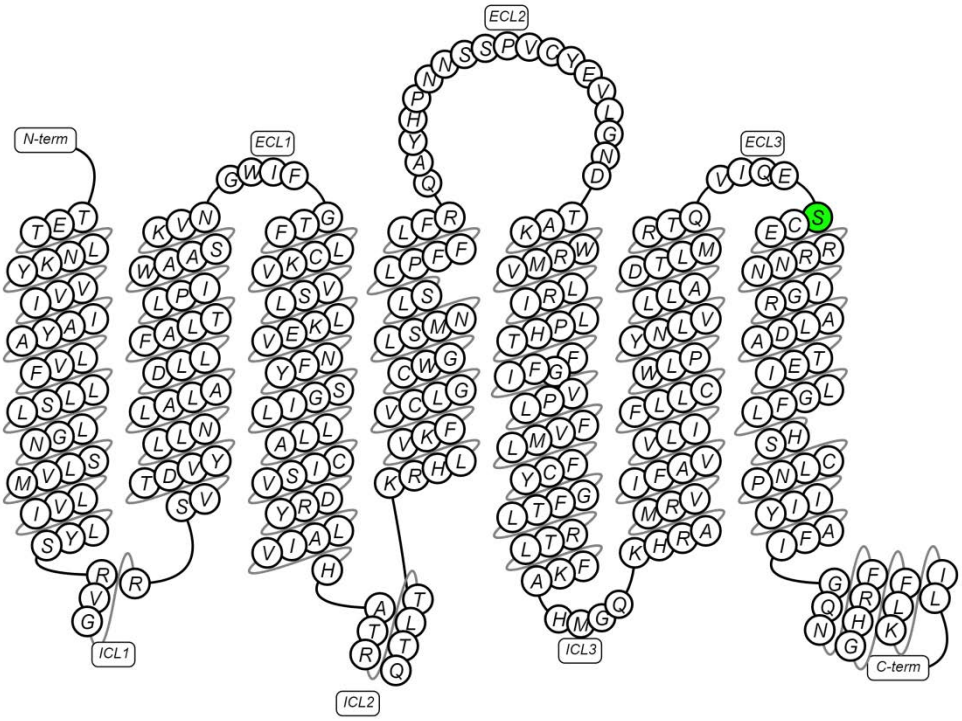


Figure 32 Two-dimensional diagram of human CXCR1 receptor and the amino acid position affected by rs2234671.

The diagram was extracted from the GPCRdb (http://gpcrdb.org/protein/cxcr1_human/). The amino acid residue affected by rs2234671 is highlighted in green.

In an evolutionary survey study comparing the sequences of CXCR1 in humans and five non-human primates, CXCR1 rs2234671 was found to be specific to humans and considered to be caused by Darwinian positive selection in an accelerated unilateral receptor evolution (Liu et al., 2005). Conventionally, the major allele and the minor allele of a non-synonymous SNP were determined according to the coding strand instead of the template strand. Most researchers interested in this SNP followed this convention and I did the same in this project. However, in dbSNP database, this SNP was reported in reverse orientation to the genome. Although annotated clearly in dbSNP database, this difference was not emphasized sufficiently in the 1000 Genomes database and some other key reference database. This has led to ambiguity in the definition of the major and minor alleles leading to

misunderstanding in the scientific community. For example, conflicting descriptions were observed in a recent publication (de Haas et al., 2014).

Unlike the rarely-investigated SULF1 rs2623047, the association of CXCR1 rs2234671 has been assessed in many diseases. The first disease that linked to CXCR1 rs2234671 was acute pyelonephritis (APN). In this study, minor alleles of both rs2234671 and rs3138060[34] were found to be associated with the risk of APN. These two SNPs were in complete LD. However, only rs3138060 was tested in a functional assay and shown to affect transcription factor binding which reduced the transcription of CXCR1 (Lundstedt et al., 2007). The C allele was also found to be associated with the susceptibility of cutaneous and mucosal leishmaniasis (OR: 1.84, 95%CI 1.04-3.23) when compared with unaffected controls (Castellucci et al., 2010). In women with asymptomatic bacteriuria, increased urine IL-8 levels was observed in C allele carriers in a gene-dose dependent manner (i.e., concentration increased by every additional C allele) (Hawn et al., 2009). In metastatic colorectal cancer patients, CXCR1 rs2234671 was shown to be predictive of bevacizumab based treatment outcome. Frequency of C allele was found to be significantly increased in patients with stable or progressive disease in comparison with those with complete or partial remission (Gerger et al., 2011). Furthermore, the C allele has been associated with the risk of active HBV infection (Almajhdi et al., 2013). In addition, Hsing and colleagues reported that the C allele carriers of this SNP had a significantly higher risk of developing gallbladder cancer among patients with gallstones (Hsing et al., 2008). However, their conclusion was drawn due to misuse of statistical methods.

34 In this study, rs2234671 and rs3138060 were named as variant 2 and variant 1 respectively.

After re-examining and re-analysing their raw data, I was unable to find any significant associations.

Only one published study has investigated the role of this SNP in HPV-related diseases and the authors failed to find an association between this polymorphism and the susceptibility to CIN3 or cervical cancer (Wang et al., 2009a). Contrary to this, the data from this project suggests that the C allele was significantly enriched in HPV16-positive cervical cancer. My colleague Dr Michelle Etherson also observed the enrichment of the C allele in HR-HPV positive vulval cancers, especially in HPV16 positive vulval cancers. The pooled results were shown in **Section 4.4.2**. The contradictory results between the published study and this study may be due to the ethnic difference and the difference of the grouping strategy. Their study was performed in a set of samples from the Costa Rican population with a minor allele frequency of 18% in the controls and 16% in HPV persistence and CIN3+ cases (Wang et al., 2009a). In our studies, the minor allele frequency in Scottish controls was 6.8%, which is close to that of the British population (8%) and East Asian population (7%) but much lower than the Global minor allele frequency (14%) according to the 1000 Genomes project (Consortium, 2015). More studies are needed to assess the role of this SNP in MAF-similar populations since the association identified might be ethnic-specific. Besides, our results showed that the allele and genotype frequencies were significantly different between HSILs and cancers. However, in the published study, CIN3 was grouped together with cervical cancer cases, which might hinder the ability to detect the underlying association specific to the cancer development. Moreover, it is noteworthy that significant interactions between HPV16 and this polymorphism were observed in the progression process

from high grade squamous intraepithelial lesions to the cancer, but not in other transition states. This finding suggested a potential specific synergistic effect of this SNP and HPV16 on cancer development. Intriguingly, a study performed by Almajhdi et al also showed a joint effect of active HBV infection and this SNP on the development of cirrhosis and hepatocellular carcinoma (Almajhdi et al., 2013). Since the biological role of CXCR1 rs2234671 has rarely been investigated, this SNP was chosen for subsequent functional assays.

Chapter 5 Functional effects of CXCR1 rs2234671

5.1 Introduction

As described previously, it is always difficult to investigate the functional roles of a given SNP even if it is truly functional. Phenotypic changes caused by most SNPs are almost never as obvious as those caused by rare mutations in monogenic disorders inherited according to Mendelian genetics. In particular, non-synonymous SNPs are more likely to fine–tune the functions of a gene rather than completely abrogate or re-shape its original functions. Therefore, it is necessary to have a more concrete understanding of the pathway in which the SNP-containing gene was involved. The CXCR1 SNP rs2234671 further investigated in this project is involved in the chemokine signalling pathway IL-8-CXCR1.

Since discovered and purified by Yoshimura and colleagues in 1987 (Yoshimura et al., 1987), the functional role of IL-8 has been extensively studied in different physiological and pathological conditions. As the accepted prototypic chemokine of the ELR+ CXC chemokine family, IL-8 is characterized by the glutamic acid-leucine-arginine motif (ELR motif) immediately preceding the first cysteine of the conserved CXC motif[35]. It has been shown to mediate its effects through two receptors, CXCR1 and CXCR2 (Griffith et al., 2014). Both receptors belong to the GPCR superfamily and have been found to be naturally expressed on a variety of cells, including neutrophils (Stillie et al., 2009), natural killer cells (Lima et al., 2015), endothelial cells (Li et al., 2003; Li et al., 2002; Singh et al., 2011), epithelial cells (Palter et al., 2001), mast cells (Inamura et al., 2002) and bone marrow-derived mesenchymal stem cells (Park et al., 2015; Ringe et al., 2007). Although these two

35 CXC motif: the first two adjacent cysteines of the four conserved cysteine residues near the N-terminus are separated by a random single amino acid.

receptors are frequently co-expressed and share high similarity (75.3%[36] amino acid identity), there is plenty of evidence showing that they are not completely functionally redundant.

The ligand selectivity of the two receptors is distinctly different. CXCR1 has only been consistently shown to bind IL-8 (or called CXCL8) and CXCL6 with high affinity (Holmes et al., 1991; Wuyts et al., 1998). A recent study also demonstrated an essential role for CXCR1 in CXCL3 induced airway smooth muscle cell migration (Al-Alwan et al., 2013). Unlike the highly selective CXCR1, CXCR2 serves as a relatively promiscuous receptor, which responds to nearly all ELR+ CXC chemokines and even some of their orthologs in other species (Gilli et al., 2005; Griffith et al., 2014). Although both receptors bind to IL-8 with comparable affinity (Jones et al., 1997; Lowman et al., 1996), whether they act similarly or differently depends upon the intracellular signalling machinery they are coupled to, and more importantly, the biological condition and the cell types they are found in. Physiologically, neutrophils are the main cell type expressing both CXCR1 and CXCR2. In neutrophils, intracellular Ca^{2+} mobilization, actin reorganization, chemotaxis, exocytosis and suppression of apoptosis have been shown to be mediated by both receptors, while superoxide production and the activation of phospholipase D are only mediated through CXCR1 and the release of MMP-9 exclusively depends on CXCR2 (Aul et al., 2012; Chakrabarti and Patel, 2005; Glynn et al., 2002; Jones et al., 1997; Jones et al., 1996; Kredel et al., 2014). Hartl and colleagues also demonstrated that both the oxygen-dependent and oxygen-independent bacterial killing function of neutrophils are solely facilitated by

36 Obtained by comparing NCBI Reference Sequence NP_001548.1 and NP_000625.1 using EMBOSS Needle

CXCR1 (Hartl et al., 2007). Endothelial cells are another cell type that exhibit widespread expression of IL-8 receptors but there are inconsistent results in the literature in terms of their roles in endothelial cell proliferation and migration. The discrepancies are likely due to the different sources of the cells and different culture methods (Li et al., 2005; Okada et al., 2009; Schraufstatter et al., 2001). More recently, both CXCR1 and CXCR2 were found to be involved in promoting the survival, proliferation, migration and invasion of endothelial cells in response to IL-8. Additionally, they are also responsible for the induction of adhesion molecules and the formation of filopodia, microvilli and capillary-like structure (Singh et al., 2011; Whittall et al., 2013). In airway smooth muscle cells, both receptors were found to be responsible for IL-8 induced cell contraction, migration and increase of intracellular $Ca2^{+}$ (Govindaraju et al., 2006). In contrast, bladder urothelial cells bear both IL-8 receptors but only CXCR1 could functionally support cell survival and proliferation following IL-8 stimulation (Tseng-Rogenski and Liebert, 2009). Gingival keratinocytes also have both IL-8 receptors but each has a different pattern of intracellular distribution (Sfakianakis et al., 2002). Apart from cells that co-express both IL-8 receptors, there are many cell types reported to express only one of the receptors. For example, CXCR1, but not CXCR2, was detected on CD8+ cytolytic effector T cells, regulatory T cells and endometrial stromal cells and was shown to mediate migration or proliferation of these cells in response to IL-8 (Eikawa et al., 2010; Li et al., 2012; Takata et al., 2004).

CXCR1 and CXCR2 can be activated *in vitro* by IL-8 at relatively low concentration, usually 10-50ng/ml. Much higher concentration of IL-8 (e.g.1000ng/ml) can lead to the internalization and desensitization of both receptors (Feniger-Barish et al., 2003).

Compared with CXCR2, CXCR1 is less susceptible to the desensitization effect of IL-8 and its recovery occurs very fast. Therefore it was thought to act predominantly within the inflammatory site where a relatively high concentration of IL-8 was found (Chuntharapai and Kim, 1995). The different responses towards IL-8 between CXCR1 and CXCR2 were considered to be largely dependent on their amino acid differences, which are unevenly distributed across the whole sequence. Structurally, the two receptors diverge from each other mainly in four regions: the extracellular N-terminus (residues 1-39[37]), the 4th trans-membrane domain (residues 155-174), the second extracellular loop (residues 175-199) and the intracellular C-terminus (residues 209-250). All of these regions were shown to be functionally important. Previous studies demonstrated that the N terminus, the 4th trans-membrane domain and the second extracellular loop are essential for the sequential two-step ligand binding (Joseph et al., 2010; Rajagopalan and Rajarathnam, 2004). The C terminus and the second extracellular loop together were found to be involved in receptor internalization and desensitization (Nasser et al., 2007; Richardson et al., 2003). Additionally, the presence of two disulfide bonds in CXCR1, Cys30-Cys277 and Cys110-Cys187 respectively, was also shown to be important in maintaining the three-dimensional structure of the receptor and facilitating ligand binding (Park et al., 2012).

In pathological conditions, cell based models and model organisms are frequently used to explore the functional effects of each participant in a signalling pathway. However, unlike other well investigated chemokine-chemokine receptor pairs, no IL-8 ortholog has been identified in conventional rodent models. Although both

37 According to the amino acid sequence of CXCR1

mouse and rat orthologs of CXCR1 have been identified, they exhibit only 64.7% and 65.5% [38]amino acid identity to the human CXCR1 respectively. Rat CXCR1 was shown to be expressed in the lung and in macrophages but was undetectable in neutrophils with/without LPS stimulation (Dunstan et al., 1996). Similarly, mouse CXCR1 is also not a good representative for human CXCR1. Although this mouse ortholog has a similar distribution to the human counterpart, conflicting results were observed in terms of their functions induced by corresponding human and murine ligands (Fan et al., 2007; Fu et al., 2005; Moepps et al., 2006). Collectively, these data suggest that rodent orthologs were unable to fully mimic the human IL-8-CXCR1 signalling.

Due to the lack of the IL-8 orthologs in rodents, cell models were widely utilized to investigate the role of IL-8-CXCR1 signalling in disease conditions. To investigate the functional effects of CXCR1 and CXCR2 separately, selective blockade or knock-down of either receptor has often been performed on cells endogenously expressing both receptors using neutralizing antibodies, small molecule antagonists or RNA interference (Ha et al., 2017; Singh et al., 2010; Stillie et al., 2009). Cells that express neither IL-8 receptor were also used to clarify mechanisms through overexpression of either or both receptors. The most frequently used cell lines for transfection and transduction include HEK293 cells and RBL-2H3 cells. There is plenty of experimental evidence that, although not naturally bearing CXCR1 or CXCR2, these cell lines do possess the signalling machinery required for IL-8

38 Obtained by comparing NCBI Reference Sequence NP_839972.1 and with NP_000625.1 using EMBOSS Needle

induced activities (Cohen-Hillel et al., 2009; Cohen-Hillel et al., 2006; Feniger-Barish et al., 2003; Han et al., 2012).

Since CXCR1 is extensively expressed on immune cells, its role in infectious and inflammatory diseases has been widely investigated. In benign inflammatory diseases, increased expression of CXCR1 on circulating immune cells or cells at the site of inflammation was frequently observed and targeting IL-8-CXCR1 signalling by antagonists or blockers was found to be beneficial in a broad range of diseases, such as asthma, chronic obstructive pulmonary disease, ulcerative colitis, cystic fibrosis and type I diabetes (Russo et al., 2014). In infectious diseases, expression of CXCR1 in each cell type is elaborately regulated after exposure to different pathogens and any inappropriate changes may lead to detrimental outcomes. CXCR1 plays a critical role in mediating the migration of neutrophils towards the infectious site. This IL-8 induced recruitment has frequently been shown to be beneficial in infectious diseases (Ha et al., 2017; Lundstedt et al., 2007; Ragnarsdottir et al., 2008). Various strategies have been developed by pathogens so as to break or weaken this defence. For example, *Helicobacter pylorus* was shown to decrease CXCR1 expression through rapid degradation and endocytosis (Schmausser et al., 2004). Mycobacteria could reduce IL-8 secretion from alveolar epithelial cells by CXCR1 and CXCR2 mediated GRK2 expression and inhibit subsequent neutrophil recruitment (Håkansson et al., 2013). *Staphylococcus aureus* was shown to kill neutrophils and monocytes by releasing Leukotoxin ED that specifically targets either of the IL-8 receptors (Reyes-Robles et al., 2013). However, in some other cases (e.g. *Klebsiella pneumoniae* infection), recruitment of neutrophils mediated by CXCR1 may do more harm than good due to severe neutrophilic inflammation (Wei

et al., 2013). In viral infections, IL-8 triggered by adenovirus from macrophages was found to promote virus binding to epithelial cells through CXCR1 mediated αvβ3 integrin relocation (Lütschg et al., 2011). In HIV-1 infection, IL-8 was shown to enhance viral replication in T lymphocytes, monocyte-derived macrophages and microglia through CXCR1 mediated activation of the NF-κB pathway (Lane et al., 2001; Mamik and Ghorpade, 2014). Higher expression of both IL-8 and CXCR1 was found in PBMCs from antiretroviral therapy naïve children with progressive disease, compared with their long-term non-progressive counterparts (Pananghat et al., 2016). In addition, HIV-1 matrix protein p17 was found to mimic IL-8 via interacting with CXCR1 expressing cells, including monocytes, lymph node-derived lymphatic endothelial cells, umbilical vein endothelial cells (Caccuri et al., 2012; Caccuri et al., 2014; Giagulli et al., 2012).

Apart from infectious diseases and benign inflammatory diseases, IL-8-CXCR1 signalling has been studied in a wide variety of cancers, including melanoma, prostate cancer, breast cancer, gastric cancer, pancreatic cancer, colorectal cancer and lung cancer. Up-regulated CXCR1 was observed in many cancer types and was commonly linked to poor prognosis and resistance to chemotherapy (Chen et al., 2014c; David et al., 2016; Han et al., 2015). In colorectal cancer, expression of CXCR1 correlated with the metastatic potential of tumour cells and could be regulated by epithelial-mesenchymal transition (Bates et al., 2004; Li et al., 2001). In prostate cancer, CXCR1 was found to be highly expressed only by non-neuroendocrine secretory-type tumour cells and the distribution of CXCR1 became more non-apical and cytoplasmic as disease severity increased (Huang et al., 2005; Murphy et al., 2005). In melanoma cell lines, expression of CXCR1 was

decreased in cells put into suspension but increased noticeably after reattachment as a consequence of phosphorylation of Jun amino-terminal kinases (Uen et al., 2015). In general, the IL-8-CXCR1 pathway has been shown to act directly on tumour cells in both autocrine and paracrine manners and mainly exert oncogenic effects (Ha et al., 2017; Liu et al., 2016b; Waugh and Wilson, 2008). Suppression of IL-8 signalling through CXCR1 knockdown or antagonist in various types of tumour cells demonstrated that this pathway could fuel tumour growth by promoting cell cycle entry, supporting tumour proliferation, preventing spontaneous apoptosis and enhancing angiogenic activity, migration, invasion and metastasis (Chen et al., 2014c; Liu et al., 2012b; Shamaladevi et al., 2009; Singh et al., 2010; Wang et al., 2016a).

CXCR1 was also found to be enriched in cancer stem cells in many cancer types and to facilitate their functions in response to IL-8 (Chen et al., 2014c; Ginestier et al., 2010). Although only CXCR1 was detected on glioma cell lines, both CXCR1 and CXCR2 were expressed on the cancer stem cells isolated from glioblastoma tumour specimens and both were responsible for the proliferation, migration and neurosphere formation of these cells induced by IL-8 (Infanger et al., 2013; Raychaudhuri and Vogelbaum, 2011). Both receptors were also shown to be able to maintain the stem-like features of glioblastoma stem cells through phosphorylation of STAT3 but CXCR1 was dominant (Zhou et al., 2014). Intriguingly, in breast cancer, tumoursphere formation and survival of the cancer stem population were found to be solely mediated by CXCR1 rather than CXCR2 (Ginestier et al., 2010). In melanoma,

CXCR1 was involved in the maintenance of the quiescent state of stem cells in an ABCB5[39] controlled IL-1β/IL-8/CXCR1 signalling loop (Wilson et al., 2014).

Apart from the tumour cells, many other cells in the tumour environment can participate in carcinogenesis via CXCR1. For example, increased circulating myeloid-derived suppressor cells (MDSCs) correlated with tumour progression and infiltration of MDSCs into the tumour environment could suppress the anti-tumour activity of T cells. Previous studies demonstrated that CXCR1 is expressed on granulocytic MDSCs and was responsible for its recruitment by IL-8 (Alfaro et al., 2016; David et al., 2016). Also, migration of Foxp3+ CD4+ regulatory T cells, was shown to be mediated through CXCR1 (Eikawa et al., 2010). In prostate cancer, CXCR1 was highly expressed on adipose stromal cells. It was shown to facilitate the induction of α-smooth muscle actin on adipose stromal cells and consequently promote endothelial cell proliferation in a contact dependent manner (Zhang et al., 2016). Additionally, activation of CXCR1 on endothelial cells can directly promote tumour angiogenesis (David et al., 2016). In bone metastases of breast cancer, CXCR1-expressing osteoclasts and their precursors were found to be responsible for pro-osteoclastogenic activity (Kamalakar et al., 2014).

The role IL-8 has also been investigated in HPV-related disease. However, the majority of published studies concentrate on the expression of IL-8 and its potential upstream regulators, such as HPV viral proteins, seminal plasma, thymidine-derived sugars and tumour associated macrophages (Carrero et al., 2015; Ding et al., 2014; Liu et al., 2016a; Sales et al., 2012; Tabata et al., 2012). Very few studies have explored the role of IL-8 receptors in depth. In HPV-related cervical disease, as

39 ABCB5: ATP binding cassette subfamily B member 5

mentioned in **Chapter 1**, a chemokine profile study done in our group found that both IL-8 and CXCR1 were upregulated with increased disease severity (Data not shown). Both IL-8 and CXCR1 seem to affect the overall outcomes of cervical cancer patients (see Appendix 4). The estimated prognostic effect of IL-8 based on TCGA-CESC database is consistent with an early finding (Fujimoto et al., 2000).

Wu and co-workers showed that IL-8 increased proliferation and migration of cervical cancer cells *in vitro* and promoted tumour growth and metastasis in their xenograft model (Wu et al., 2013b). In another study performed by Liu and colleagues, IL-8 and CXCR1, but not CXCR2 were found to be significantly increased in cervical cancer tissues in comparison with cervicitis and these differences were postulated as the consequence of hypoxia. Furthermore, the authors found that IL-8 could enhance the viability of HeLa and SiHa cells and reduce their apoptosis (Liu et al., 2014). It is noteworthy that the expression patterns of CXCR1 and in cervical cancer tissue have not been well characterized yet.

The targeted CXCR1 rs2234671 SNP causes a missense amino acid substitution in the third extracellular loop of the CXCR1 (at amino acid position 276) and the minor allele was significantly enriched in HPV16 positive cervical cancers and vulval cancers. Based on the known roles of IL-8-CXCR1 signalling in other physiological and pathological situations and after due consideration of the identified genetic association and related findings in cervical cancer, the effect of CXCR1 rs2224671 on cell migration, proliferation and angiogenesis was further investigated. The working hypothesis was that **"The G-C substitution at nucleotide position 827 (serine-threonine substitution at amino acid position 276) of CXCR1 results in increased IL8-induced cell motility, proliferation and angiogenic activity"**. Like

previous studies using selective point mutations to locate the structural features involved in ligand binding, receptor activation and regulation, in this chapter, the hypothesis was tested using cell models generated though site directed mutagenesis, which were either transient or stably over-expressing CXCR1 827G or CXCR1 827C. The effects of CXCR1 827G or CXCR1 827C on cell migration towards IL-8 was compared using a chemotaxis assay (chemotactic gradient dependent migration) and a wound healing assay (non-gradient dependent migration). The results of the chemotaxis assay and the wound healing assay are described in **Section 5.2**. Also, to assess the functional consequence of the SNP rs2234671 on tumour cell proliferation, I measured the cell growth rate using a cell counting kit-8 (CCK8) assay with HPV16 positive cervical cancer cell lines over-expressing either CXCR1 827G or its C-substituted counterpart. The results of the proliferation assay are described in **Section 5.3**. In addition, the functional role of the SNP in angiogenesis was explored. Due to the lack of an accessible endothelial cell model, the direct effect of the SNP on endothelial cells in response to IL-8 could not be explored. Instead, the indirect effects of the SNP on the IL-8 induced production of angiogenic molecules were investigated using the established cell models. Previous studies showed that a CXCR1/2 antagonist could reduce vascular endothelial growth factor (VEGF) expression in a xenograft model for lung cancer(Khan et al., 2015). In prostate cancer cells, production of VEGFA was found to be significantly increased after transfected with an IL-8–secreting plasmid and significantly decreased after CXCR1 depletion. (Araki et al., 2007; Shamaladevi et al., 2009). Collectively, these findings suggested a potential mechanism for IL-8-CXCR1 mediated angiogenic responses in cancer cells. Therefore, expression of VEGFA following IL-8 stimulation was

chosen to be measured in cells over-expressing CXCR1 827G or CXCR1 827C. The corresponding results are described in **Section 5.4**.

5.2 Effects of CXCR1 rs223467 on IL-8 induced cell motility

5.2.1 Overexpression of CXCR1-827C but not CXCR1-827G enhances chemotaxis towards IL-8 in HEK293 cells

To assess whether the serine to threonine conversion at amino acid position 276 could affect the chemotactic function of CXCR1, a chemotaxis assay was performed in a transwell system with HEK293 cells expressing plasmids encoding for the major and minor alleles, CXCR1 827G and 827C, respectively. The detailed design of this experiment was described in **Section 2.3.6.1.1**. Representative images of migrated cells in each condition were shown in Figure 33. Chemotaxis was analyzed using two-way ANOVA and Sidak's multiple comparisons test, with IL-8 and different transfectants (i.e. pCR3, pCR3-CXCR1 827G and pCR3-CXCR1 827C) as two independent variables. As expected, a significant main effect of the CXCR1 rs2234671 genotype on chemotaxis was found, $F(2, 48) = 22.41$, $p<0.0001$. Significantly more pCR3-CXCR1 827C-expressing cells were found to migrate in 6h than pCR3-transfected cells and pCR3-CXCR1 827G transfected cells. Also the presence of a significant effect of IL-8 (100ng/ml) on chemotaxis was confirmed, $F(1, 48) = 13.71$, $p<0.001$, which indicated a significantly higher cell migration rate in IL-8 treated groups. More importantly, a significant interaction between IL-8 and CXCR1 rs2234671 genotype was observed, $F(2, 48) = 4.869$, $p=0.0119$. As can be seen from both Figure 33 and 34, the number of migrated cells did not differ among non-IL-8-treated groups, which suggested that the rate of non-specific migration was similar among groups in the assay. After incubation in the presence of IL-8 for 6h,

although very few pCR3-transfected cells and pCR3-CXCR1 827G transfected cells were observed to migrate towards the bottom chamber, cells expressing pCR3-CXCR1 827C exhibited significantly increased IL-8-dependent migration.

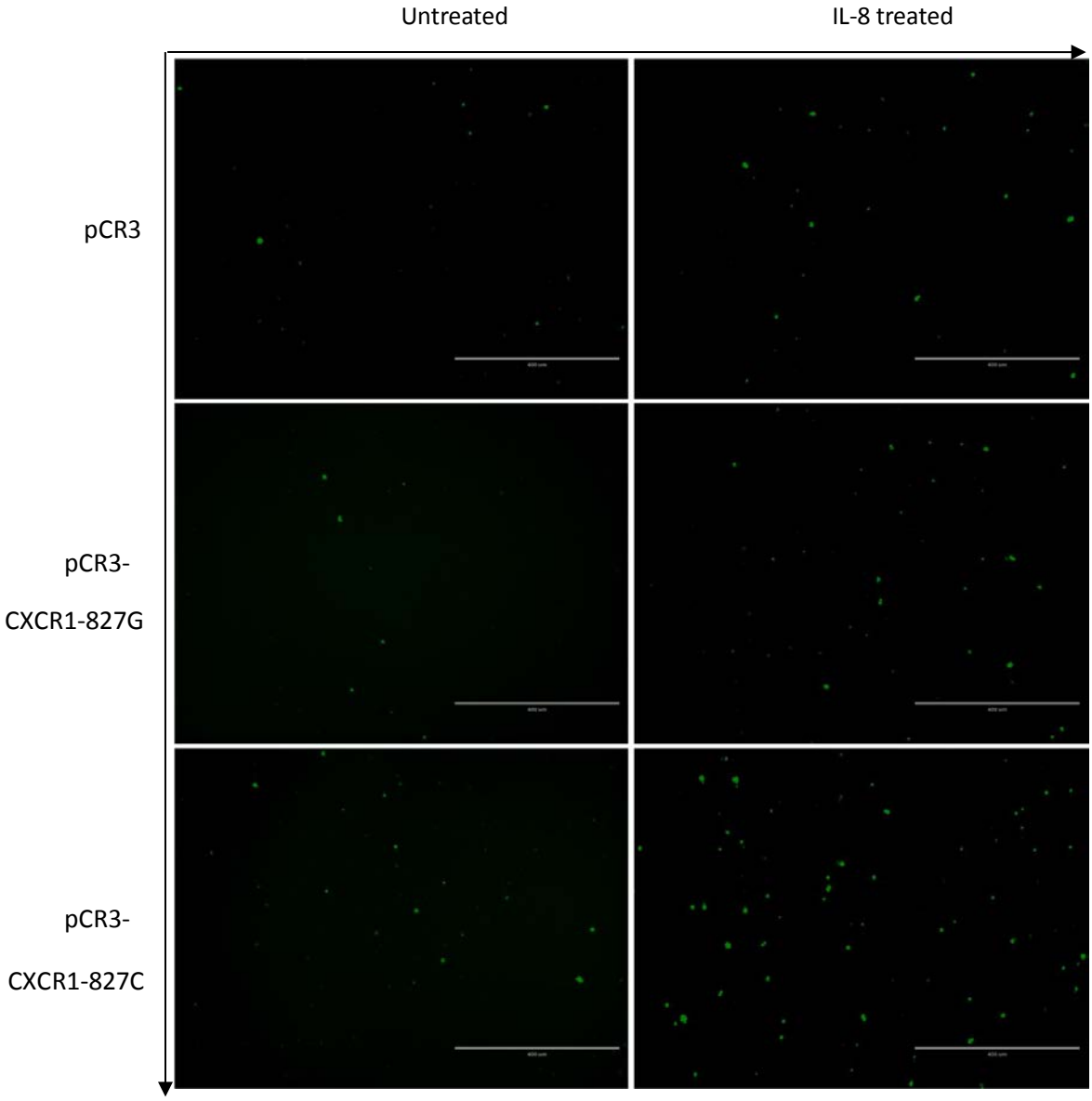


Figure 33 Chemotaxis of HEK293 cells expressing CXCR1-827G/C alleles after IL8 stimulation

This figure shows the representative images of migrated cells after 6h of incubation. Migrated cells were labelled with CFSE and shown as green dots. All these images were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) with GFP cube and the scale bar=400µm.

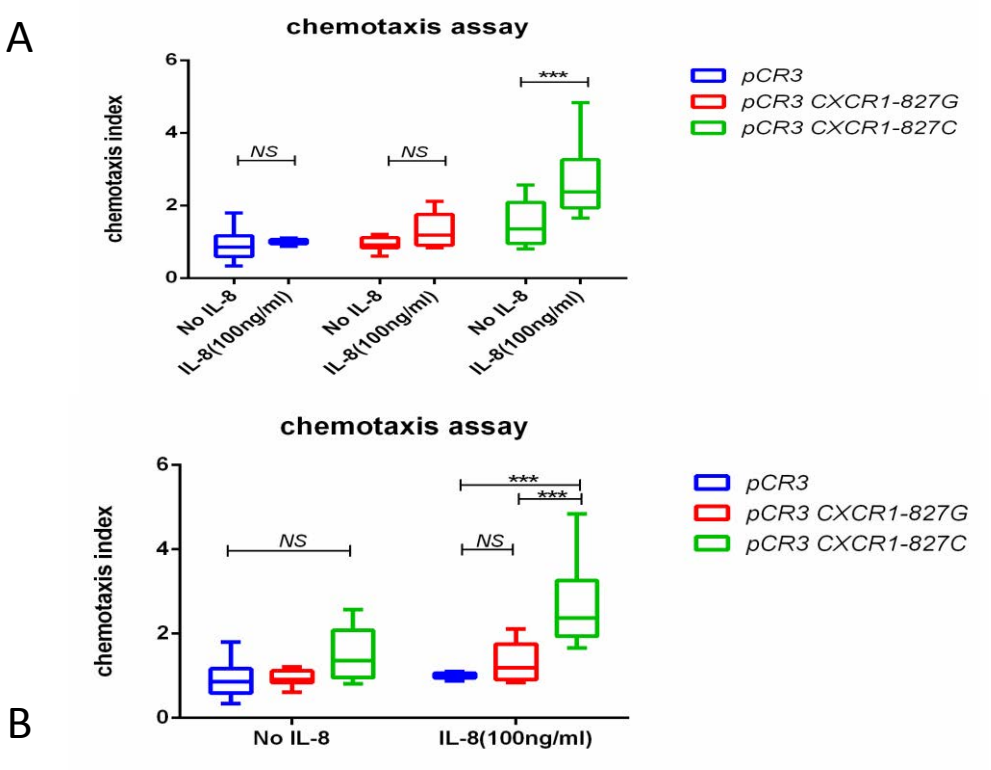


Figure 34 Chemotaxis in HEK293 cells transiently transfected with CXCR1 827G/C alleles

Chemotaxis assays were performed in a transwell system using HEK293 cells expressing pCR3, pCR3-CXCR1 827G and pCR3-CXCR1 827C.After 6h incubation in the presence or absence of IL-8(100ng/ml), cells migrated through the membrane were counted and the chemotaxis index was calculated for each transwell following the equation mentioned in Materials and Methods section. Data are presented as box-and-whiskers plots with median (horizontal line in box), the first and third quartiles (boxes) and range (whiskers) for three assays with each condition performed in triplicate. Since all groups passed normality tests, ordinary two-way ANOVA test and Sidak's multiple comparisons test were performed to test the effects of IL8 and transfectant type individually and the interaction between the two on chemotaxis. IL-8, transfectant type and the interactions between the two were all found to be significant (p value: =0.0005, <0.0001 and 0.0119 respectively). Figure 34A shows that the chemotaxis index of IL-8 treated pCR3-CXCR1 827C-expressing cells are significantly higher (Mean=2.702, SD=1.001) than their untreated counterparts (Mean=1.503, SD=0.618).No significant differences were observed for the other two groups before or after IL-8 treatment. Figure 34B indicates that the chemotaxis index of pCR3-CXCR1 827C transfected cells (Mean=2.702, SD=1.001) are significantly higher than cells transfected with pCR3-CXCR1 827G (Mean=1.324, SD=0.471) and pCR3 (Mean=1, SD=0.071) in the presence of IL-8. NS means not significant; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

5.2.2 Increased cell migration in CaSki cells over-expressing the CXCR1 827C allele

Cell-cell contacts affect cell motility. Since the chemotaxis assay mentioned above could not indicate the influence of cell-cell contacts on cell migration, a wound healing assay was performed to investigate the role of rs2234671 in cell-cell contact related cell movement in the presence or absence of exogenous IL-8 stimulation. A previous study showed that the migration of cervical cancer cells can be promoted by IL-8 (Zhao et al., 2013). Considering that rs2234671 was found to be associated with the development of HPV16 positive cervical cancers, CaSki, a HPV16 positive cell line, was chosen to be transduced with CXCR1 827G or CXCR1 827C and used in the wound healing assay. This cell line has often been utilized in wound healing assays and forms a good quality monolayer. The details of the experiment were described in **Section 2.3.6.1.2**. Flow cytometry indicated that CXCR1 expression of the tested CXCR1 827G and CXCR1 827C transduced cells were generally comparable, whereas the MFI was slightly lower in CXCR1 827C transduced cells in comparison to their CXCR1 827G counterparts (Figure 24).

The wound healing assay was first performed in culture media with 10% fetal bovine serum (FBS). Representative images of wound closure in each condition are shown in Figure 35 (in the presence of IL-8) and Figure 36 (in the absence of IL-8). Cell migration was evaluated as reduced area of the wound after 6h of incubation in the presence or absence of IL-8. Similarly to the analysis for chemotaxis, two-way ANOVA test and Sidak's multiple comparisons test were used to analyse the data, with IL8 treatment and CaSki cells with different transduction procedures (i.e.

parental (untransduced), CXCR1 827G transduced and CXCR1 827C transduced) as two independent variables.

Again, main effects of both independent variables were observed. Significantly larger average reduced area was found in IL-8 treated groups compared with those untreated ones, with a p value of 0.0123, F (1, 48)=6.775. This main effect was mainly contributed by CXCR1 827C transduced cells. No significant differences were observed before or after IL-8 treatment for the CXCR1 827G transduced and parental CaSki cells, which was similar to the findings for the pCR3-CXCR1 827G and pCR3 transfected HEK293 cells in the chemotaxis assay. Additionally, the influences of parental, CXCR1 827G and CXCR1 827C transduced CaSki cells on cell migration were significantly different, $p<0.0001$, F (2, 48) =178.6. The interaction between IL-8 and different transductants were also found to be significant, with F (2, 48) = 4.779 and p =0.0128, even though it contributed to only 2.3% of overall variability. As shown in Figure 37, CXCR1 827C transduced cells consistently migrated more than CXCR1 827G transduced cells in both IL-8 treated and untreated conditions, though the relative difference was more pronounced in the presence of IL-8. The results suggested that rs2234671 did affect CaSki cell migration in a cell-cell contact condition and the effect depended only slightly on exogenous IL-8 stimulation. It is worthwhile to note that even without exogenous IL-8, significantly more cells migrated towards the wound region in CXCR1 827G or CXCR1 827C transduced groups, compared with the parental cells. This could possibly be explained by endogenous IL-8 produced by CaSki cells. An early study showed that local production of IL-8 is significantly higher in cervical cancer patients than those with benign lesions (Tjiong et al., 1999). Previous results in the

group of my PhD supervisor suggested that CaSki cells express a higher level of IL-8 mRNA than other cervical cancer cell lines tested, including SiHa, HeLa and C-33A cells. This speculation could be experimentally tested by measuring the secreted IL-8 in the culture media or by targeting IL-8 using neutralizing antibodies.

Hypoxia and nutrient deprivation often occur at the tumour site. In order to evaluate the influence of rs2234671 on cell-cell contact related cell movement in a serum nutrient-reduced environment and diminish the wound closure caused by cell proliferation, the wound healing assay was also performed in 2% serum culture medium ("serum-starved cultures"). Medium containing less than 2% serum was tested but could not support the formation of a good quality monolayer and the creation of the wound region. All other experimental conditions remained the same. Representative images of wound healing in serum-starved condition are shown in Figure 38 (in the presence of IL-8) and Figure 39 (in the absence of IL-8). In general, the findings in 2% serum condition were similar to the 10% serum condition, though the reduced area largely decreased in all the groups and the differences among the groups became more obvious. As can be seen from Figure 40, both independent variables were shown to be significant and the interaction between the two was also found to be significant, $F(2, 48) = 5.539$, $p = 0.0068$, though it was consistent with the patterns seen in the main effects. Only CXCR1 827C transduced CaSki cells exhibited significantly greater migration activity after stimulation with exogenous IL-8, which was consistent with the observations in 10% FBS. When the cells were incubated in the absence of exogenous IL-8, the average reduced area was significantly larger for CXCR1 827C transduced cells ($M = 110394.3$, $SD = 13778.37$) than that of CXCR1 827G ($M = 78978.56$, $SD = 14082.2$) and parental CaSki cells

(M=40194.39, SD=9807.423). Likewise, when the cells were treated with IL-8, the closure area was also significantly larger for CXCR1 827C transduced cells (M=143823.8, SD=1455.070) than for CXCR1 827G (M=78978.56, SD=14082.2) and parental CaSki cells, but the difference between CXCR1 827C and CXCR1 827G transduced CaSki cells became more pronounced.

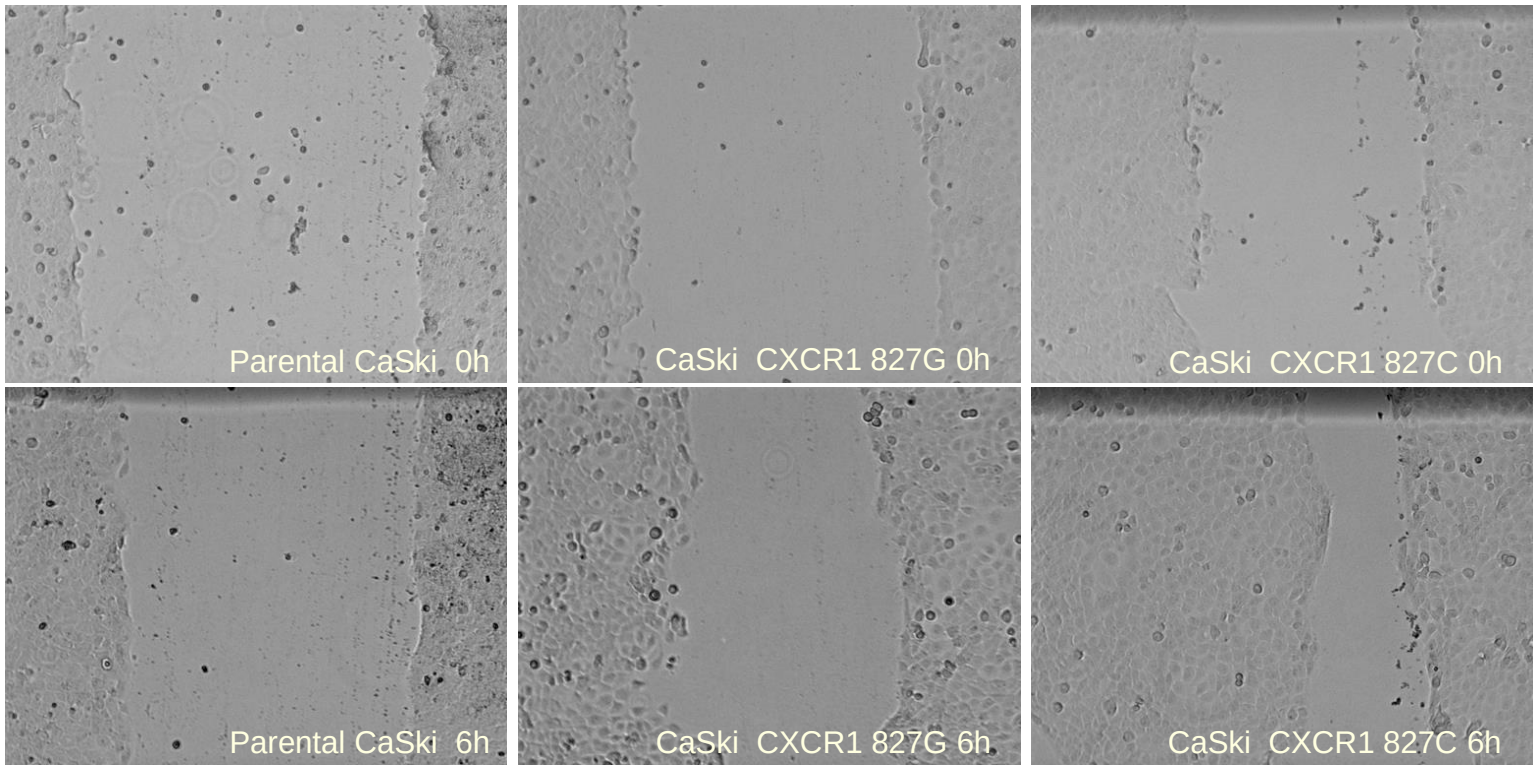


Figure 35 Representative images of cells migrating towards the wound before and after 6h of incubation in normal medium with IL-8.

All these images were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) with the transmitted filter cube. The wound area of each image was circled and measured using the freehand selection tool and the analyze-measure tool of ImageJ. Reduced area covered by migrated cells after incubation was calculated by subtracting the wound area at 6h from the wound area at 0h.

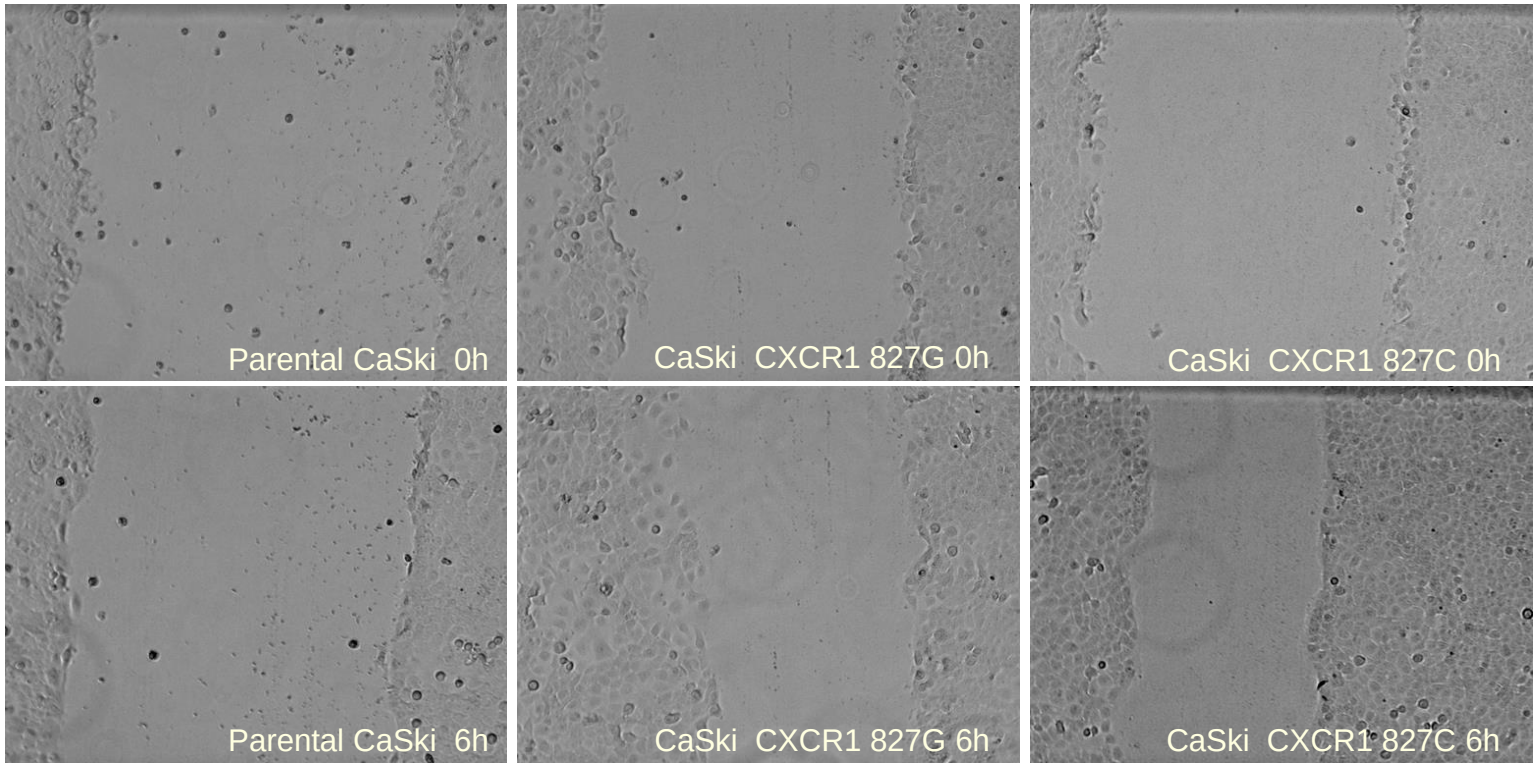
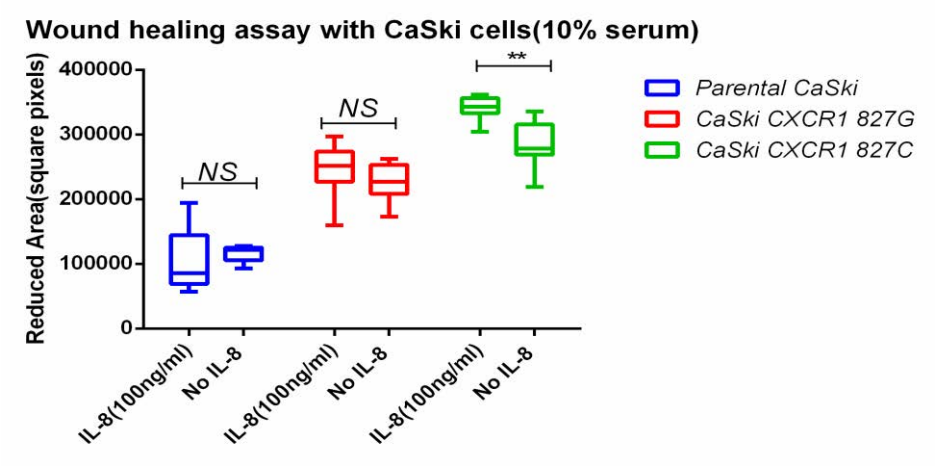


Figure 36 Representative images of cells migrating towards the wound before and after 6h of incubation in normal medium without exogenous IL-8

All these images were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) with the transmitted filter cube. The wound area of each image was circled and measured using the freehand selection tool and the analyze-measure tool of ImageJ. Reduced area covered by migrated cells after incubation was calculated by subtracting the wound area at 6h from the wound area at 0h.

A



B

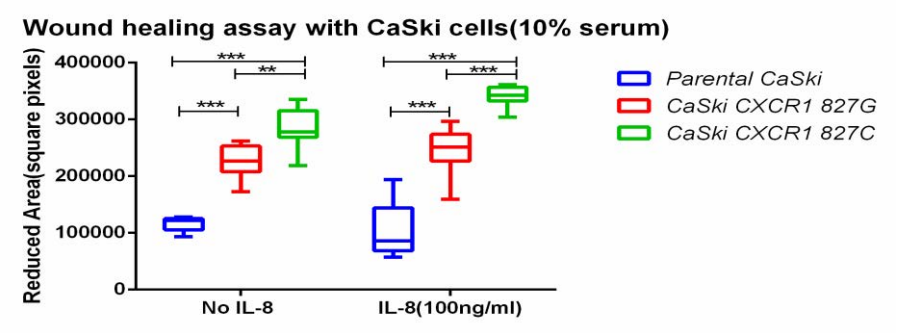


Figure 37 Migration of the CXCR1 827G transduced, CXCR1 827C transduced and parental CaSki cells in the wound healing assay (10% serum)

Wound healing assays were performed in 24 well plates using parental CaSki cells and CaSki cells stably transduced with CXCR1 827G or CXCR1 827C. After 6h incubation in the presence or absence of IL-8 (100ng/ml), reduced wound area covered by cells migrating towards the wound was measured as described in Materials and Methods section. Data are presented as box-and-whiskers plots with median (horizontal line in box), the first and third quartiles (boxes) and range (whiskers) for three assays with each condition performed in triplicate. Since all groups passed normality tests, ordinary two-way ANOVA test and Sidak's multiple comparisons test were performed to test the effects of IL8 and transductant type individually and the interaction between the two on cell-cell contact related migration. IL-8, transductant type and the interactions between the two were all found to be significant (p value: =0.0123, <0.0001 and 0.0128 respectively). Figure 37A shows that the reduced area of IL-8 treated CXCR1 827C transduced cells is significantly larger (Mean=341404.8, SD=17460.52) than their untreated counterparts (Mean=283902, SD=35043.66). No significant differences were observed for the other two groups before or after IL-8 treatment. Figure 37B indicates that the reduced area of CXCR1 827C transduced cells is significantly larger than cells transduced with CXCR1 827G both in the presence (341404.8±17460.52 vs. 245944.3±39876.680) and in the absence of IL-8(283902±35043.66 vs. 225854.3±28789.79). NS means not significant; * P < 0.05; **P < 0.01; ***P < 0.001.

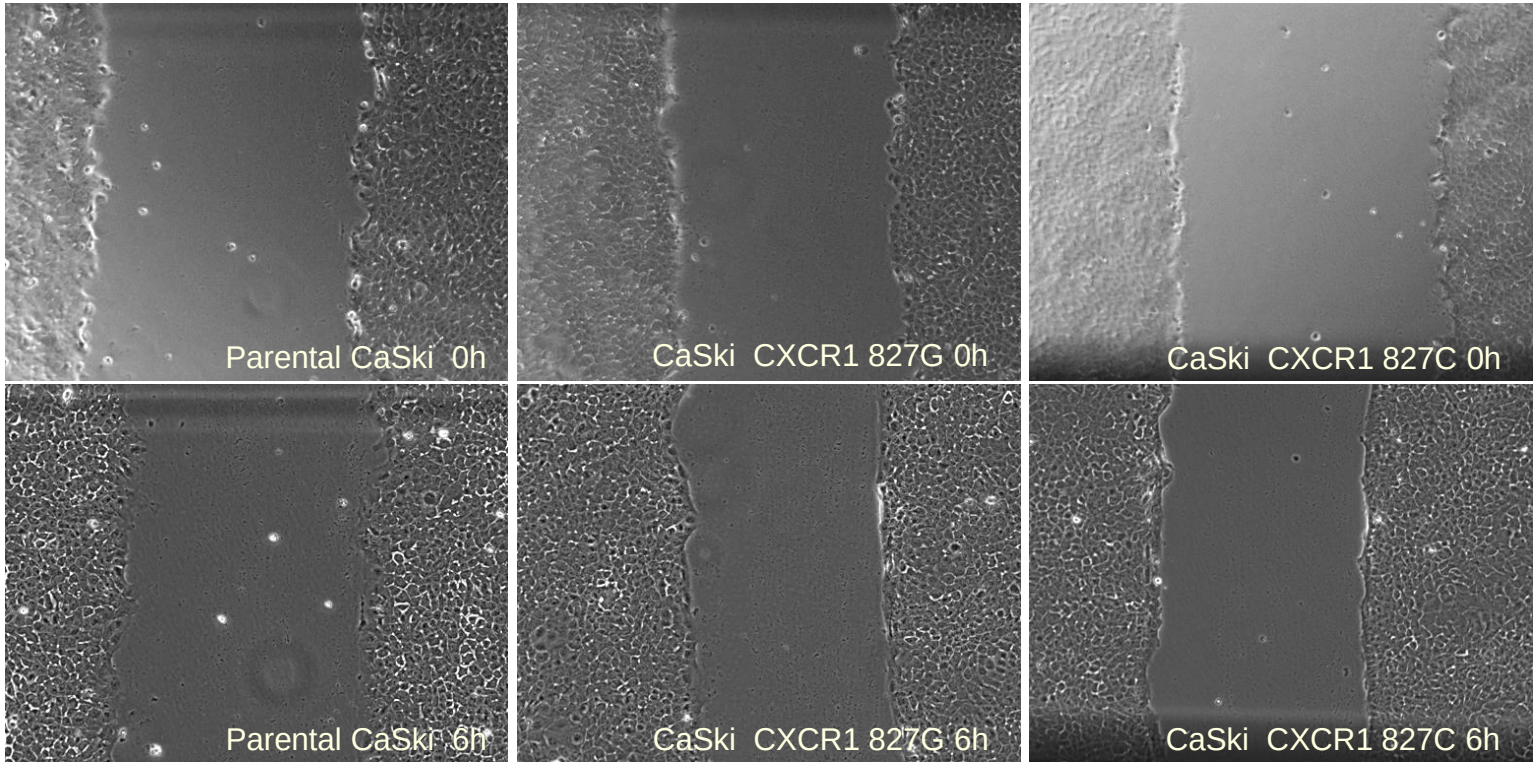


Figure 38 Representative images of cells migrating towards the wound before and after 6h of incubation in 2% serum medium with IL-8

All these images were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) with the transmitted filter cube. The wound area of each image was circled and measured using the freehand selection tool and the analyze-measure tool of ImageJ. Reduced area covered by migrated cells after incubation was calculated by subtracting the wound area at 6h from the wound area at 0h.

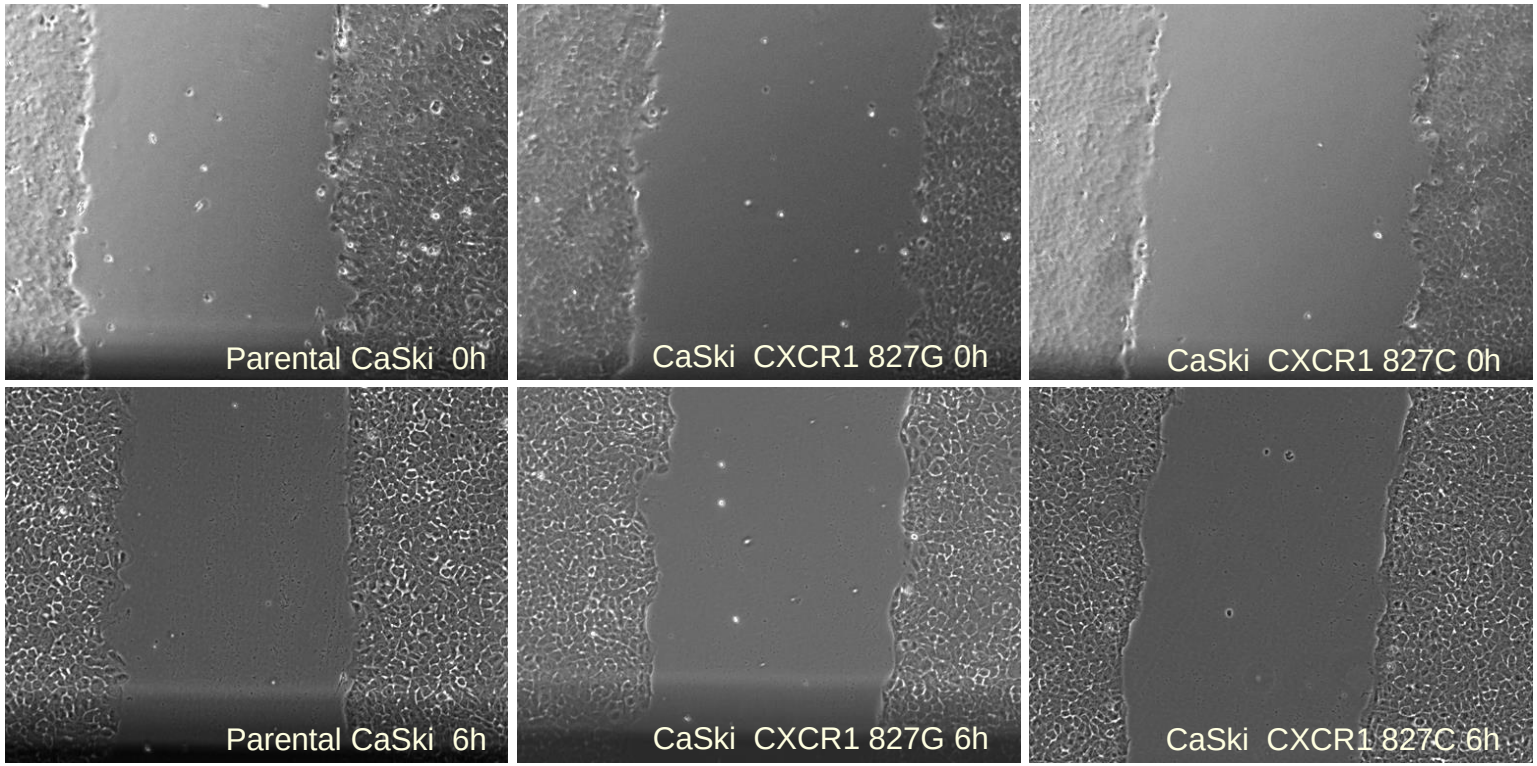


Figure 39 Representative images of cells migrating towards the wound before and after 6h of incubation in 2% serum medium without exogenous IL-8

All these images were taken at 10X under an EVOS™ digital inverted microscope (Advanced Microscopy Group, Bothell, WA) with the transmitted filter cube. The wound area of each image was circled and measured using the freehand selection tool and the analyze-measure tool of ImageJ. Reduced area covered by migrated cells after incubation was calculated by subtracting the wound area at 6h from the wound area at 0h

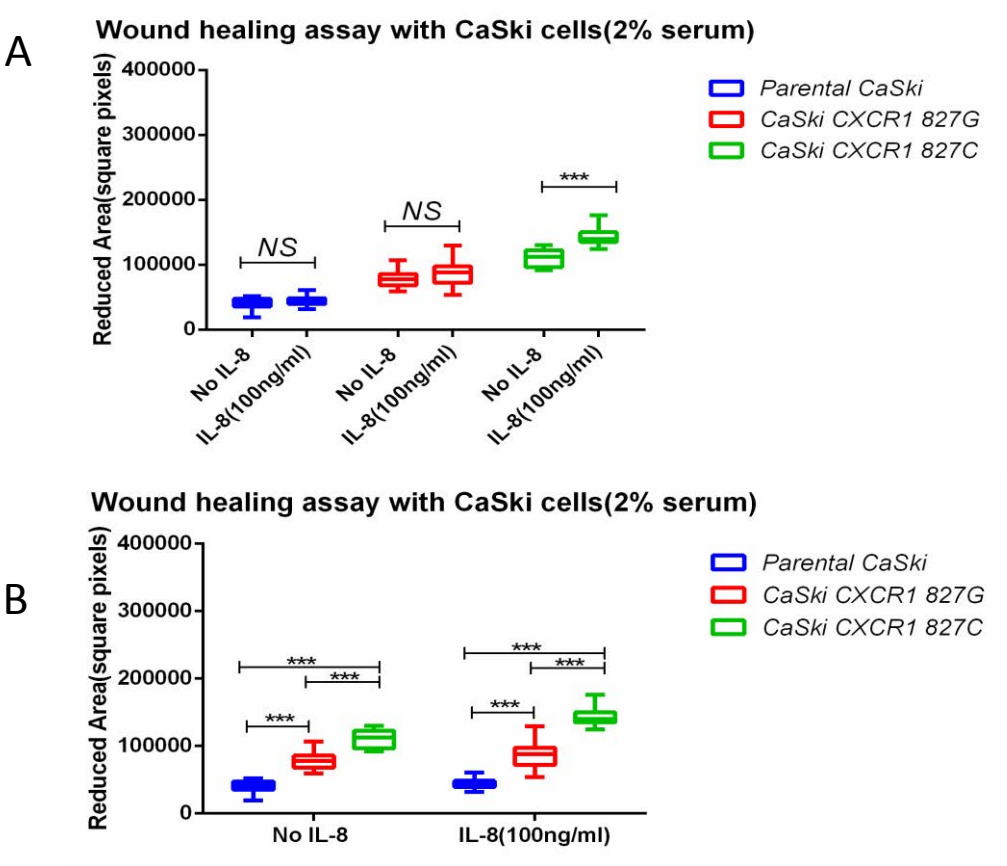


Figure 40 Migration of the CXCR1 827G transduced, CXCR1 827C transduced and parental CaSki cells in the wound healing assay (2% serum)

Wound healing assays were performed in 24 well plates using parental CaSki cells and CaSki cells stably transduced with CXCR1 827G or CXCR1 827C. After 6h incubation in the presence or absence of IL-8 (100ng/ml), reduced wound area covered by cells migrating towards the wound was measured as described in Materials and Methods section. Data are presented as box-and-whiskers plots with median (horizontal line in box), the first and third quartiles (boxes) and range (whiskers) for three assays with each condition performed in triplicate. Since all groups passed normality tests, ordinary two-way ANOVA test and Sidak's multiple comparisons test were performed to test the effects of IL8 and transductant type individually and the interaction between the two on cell-cell contact related migration. IL-8, transductant type and the interactions between the two were all found to be significant (p value: =0.0003, <0.0001 and 0.0068 respectively). Figure 40A shows that the reduced area of IL-8 treated CXCR1 827C transduced cells is significantly larger (Mean=143823.8, SD=14553.07) than their untreated counterparts (Mean=110394.3, SD=13778.37). No significant differences were observed for the other two groups before or after IL-8 treatment. Figure 40B indicates that the reduced area of CXCR1 827C transduced cells is significantly larger than cells transduced with CXCR1 827G both in the presence (143823.8±14553.07 vs. 87210.61±21568.40) and in the absence of IL-8(110394.3±13778.31 vs. 78978.56±14082.2). NS means not significant; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

5.3 Effects of CXCR1 rs223467 on IL-8 induced proliferation

CXCR1 is consistently reported to induce cell proliferation in response to IL-8. As mentioned in **Section 5.1**, IL-8 can promote the proliferation of both HeLa and SiHa cells (Liu et al., 2014). To assess whether rs2234671 could affect cell proliferation mediated by CXCR1 in response to IL-8 stimulation, proliferation assays were performed with stably transduced CaSki and SiHa cells over expressing either CXCR1 827G or 827C. Both these cervical cancer cell lines are HPV16 positive, which was considered to be an important feature according to the genetic association studies. Similar to the transduced CaSki cells, flow cytometry showed that CXCR1 expression of the tested SiHa transductants was generally comparable, although the MFI of CXCR1 827C transduced cells was also slightly lower than those with CXCR1 827G (Figure 23). As mentioned in **Section 2.3.6.2**, cells were seeded and treated with concentrations of IL-8, ranging from 0.1 to 100ng for 48h and then evaluated for the proliferation rate in comparison with untreated controls.

As shown in Figure 41 and 42, rs2234671 could affect cell proliferation in both CXCR1 827C transduced CaSki and SiHa cells, though the mode of action was a slightly different. CXCR1 827C transduced CaSki cells exhibited significantly higher proliferation rate than the untreated controls in response to 1ng/ml, 10ng/ml or 100ng/ml exogenous IL-8 while significant increase of cell growth was observed for CXCR1 827C transduced SiHa cells only in the presence of 10ng/ml exogenous IL-8. Cell proliferation rates of CXCR1 827G transduced CaSki cells and parental CaSki did not significantly increased compared with corresponding untreated controls. Similarly, exogenous IL-8 did not promote proliferation of CXCR1 827G transduced and parental SiHa cells.

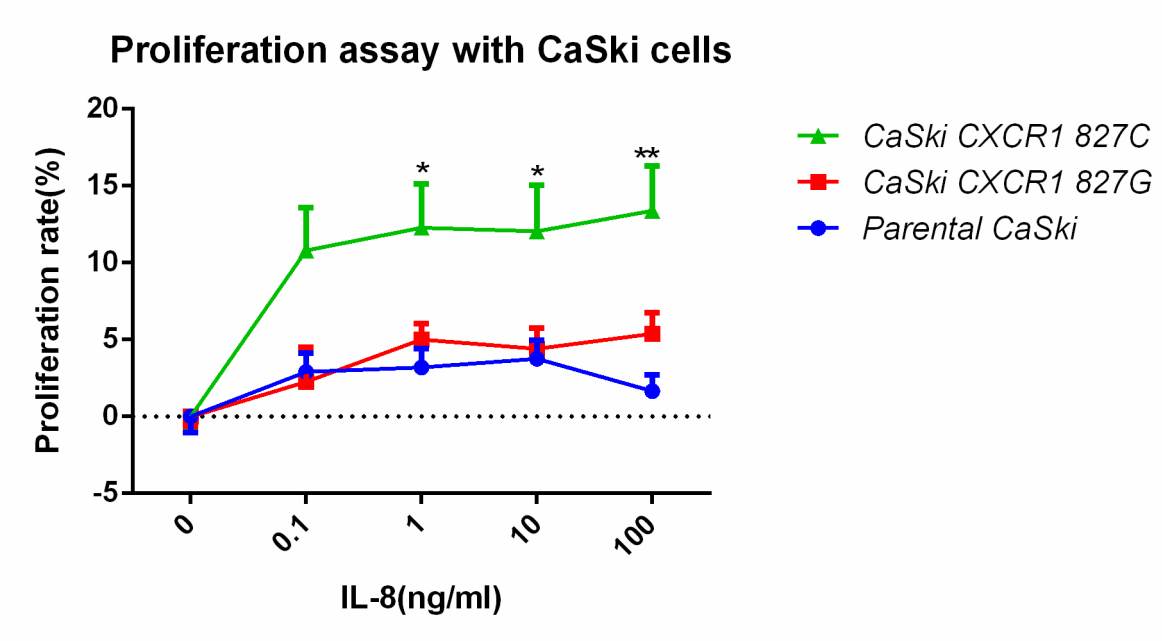


Figure 41 Proliferation of CXCR1 827G, CXCR1 827C and parental CaSki cells in the presence of different concentration of exogenous IL-8

Proliferation assay was performed to analyze the proliferation of CXCR1 827G, CXCR1 827C transduced SiHa cells and parental SiHa cells. Cells were seeded in 96-well plates and treated with different concentration of IL-8, ranging from 0.1-100ng/ml. 48h after incubation, proliferation was measured using CCK-8 cell counting kit as the increase in the absorbance and the proliferation rate was calculated compared with corresponding non-treated controls. The assay was performed three times with each condition in triplicate. Data are expressed as the mean ± SE. Since all groups passed normality tests, one-way ANOVA test and Sidak's multiple comparisons test were performed. * $P < 0.05$; **$P < 0.01$.

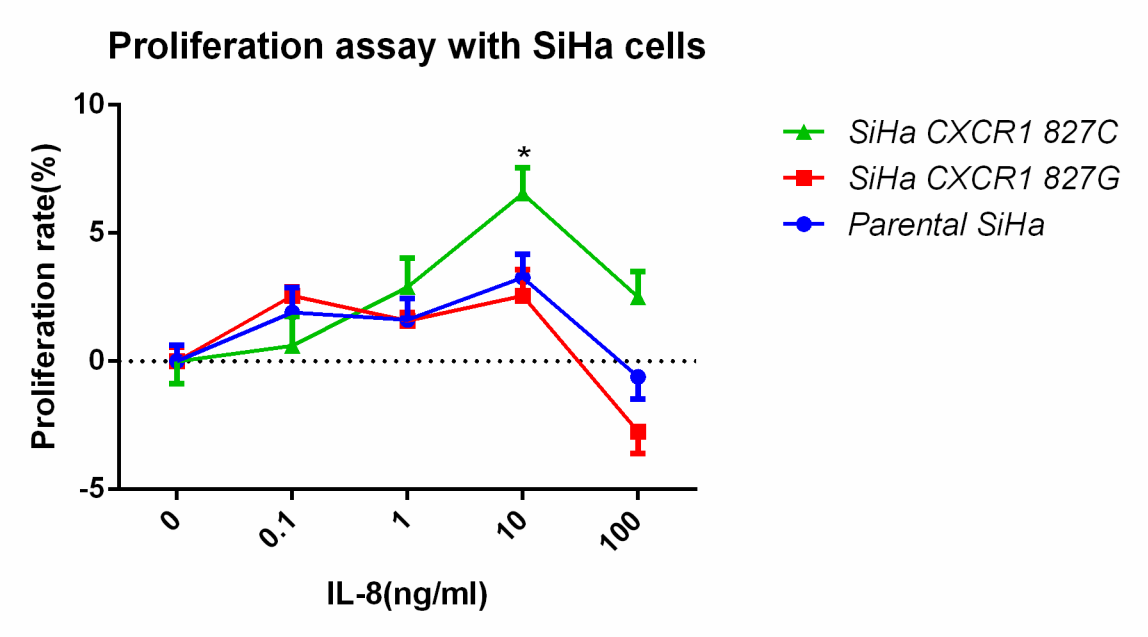


Figure 42 Proliferation of CXCR1 827G, CXCR1 827C and parental SiHa cells in the presence of different concentration of exogenous IL-8

Proliferation assay was performed to analyze the proliferation of CXCR1 827G, CXCR1 827C transduced SiHa cells and parental SiHa cells. Cells were seeded in 96-well plates and treated with different concentration of IL-8, ranging from 0.1-100ng/ml. 48h after incubation, proliferation was measured using CCK-8 cell counting kit as the increase in the absorbance and the proliferation rate was calculated compared with corresponding non-treated controls. The assay was performed three times with each condition in triplicate. Data are expressed as the mean ± SE. Since all groups passed normality tests, one-way ANOVA test and Sidak's multiple comparisons test were performed. When treated with 10ng/ml IL-8, a promoting effect on cell proliferation of CXCR1 827C transduced cells was observed, of which the proliferation rate was significantly higher than untreated controls. When treated with 100ng/ml IL-8, proliferation rates of CXCR1 827G transduced and parental SiHa cell groups significantly decreased in comparison with their corresponding 10ng/ml IL-8 treated counterparts (p value<0.001 and =0.024 respectively). No significant difference was observed between 10ng/ml and 100ng/ml IL-8 treated CXCR1 827 transduced SiHa cell groups. * $P < 0.05$.

5.4 Effects of CXCR1 rs223467 on angiogenesis indirectly induced by IL-8

5.4.1 Measuring the induction of VEGFA by IL-8 using the transient transfection model

Transiently transfected cell models were initially used for the measurement of VEGFA expression in response to IL-8 treatment (100ng/ml). pCR3, pCR3-CXCR1 827G, pCR3-CXCR1 827C transfected HEK293 cells, together with mock transfected HEK293 cells, were incubated with 100ng/ml IL-8 for a range of time periods (i.e. 6h, 8h, 12h and 24h) afterwards, VEGFA mRNA was measured for each condition as described in **Section 2.3.6.3.1**. As can be seen from Figure 43, significant differences among tested groups were observed at both 6h and 8h, but were more pronounced at 8h. After incubation with IL-8 for 8h, VEGFA expression was significantly increased in all the transfected groups in comparison with the mock-transfected group. However, no significant differences were observed among the transfected groups. The induction of VEGFA did not differ between pCR3-CXCR1 827C transfected cells and their pCR3-CXCR1 827G counterparts. Since the pCR3 transfected group also significantly increased VEGFA expression, the induction of VEGFA observed in the assay was very unlikely to be mediated by the overexpression of CXCR1 in transfected HEK293 cells.

In order to determine the factors that caused the induction of VEGFA expression in the transient transfection model, I repeated the same assay using the 8h time point, with two new groups added in, which were non-IL-8 treated pCR3 transfected HEK293 group and non-IL-8 treated mock transfected group respectively. The results were shown in Figure 44. Again no significant difference in VEGFA

expression was observed between pCR3-CXCR1 827C and pCR3-CXCR1 827C transfected cells. As shown in Figure 44, neither transfection with pCR3 in the absence of IL-8 nor mock-transfection with IL-8 exhibits any significant enhancing effect. These findings suggested that the induction of VEGFA expression in the assay was likely due to a synergistic effect of transfection with pCR3 and IL-8 treatment.

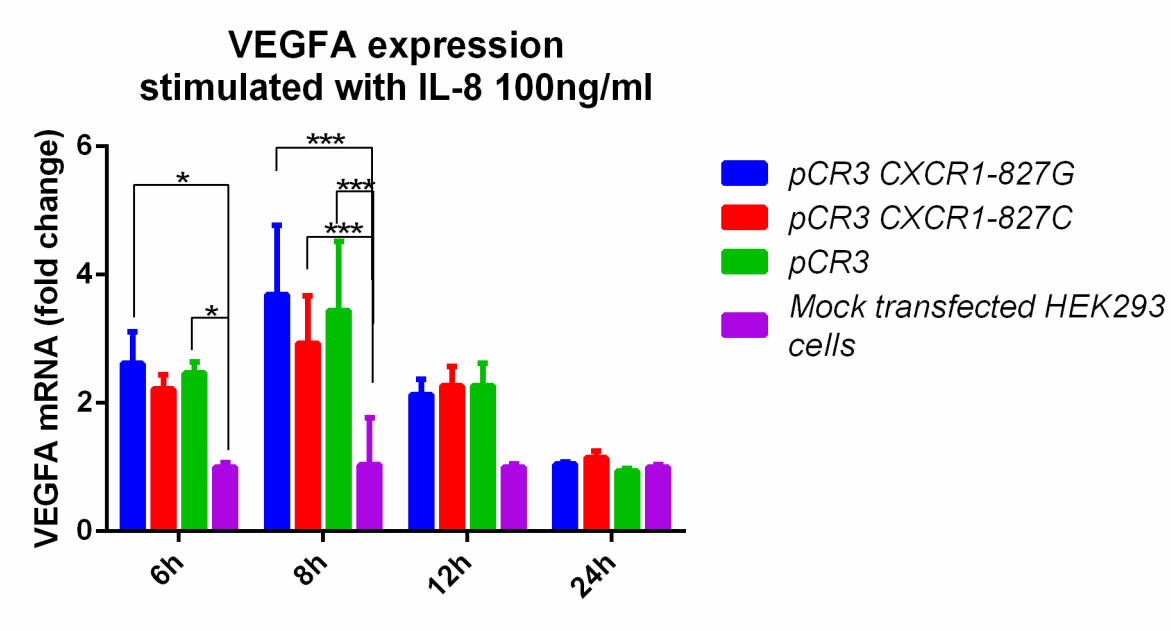


Figure 43 Time course analysis of VEGFA expression in HEK293 expressing CXCR1 827G/C alleles after IL-8 stimulation.

Expression of VEGFA mRNA extracted from pCR3, pCR3 CXCR1 827G, pCR3 CXCR1 827C transfected and mock-transfected HEK293 cells after incubation with IL-8(100ng/ml) for 6h, 8h, 12h and 24 respectively, was determined by RT-qPCR. Three independent assays were performed to obtain the data and in each assay, qPCR was performed in duplicate for each sample. GAPDH was used as housekeeping controls and the fold change of VEGFA in each test sample compared with IL-8 treated mock transfected cells was calculated by 2-ΔΔCT method as described in **Section 2.3.6.3.1**. Statistical analysis was performed using ordinary two-way ANOVA test and Sidak's multiple comparisons test. Data are presented as mean ±SD, *P<0.05, ***P<0.001

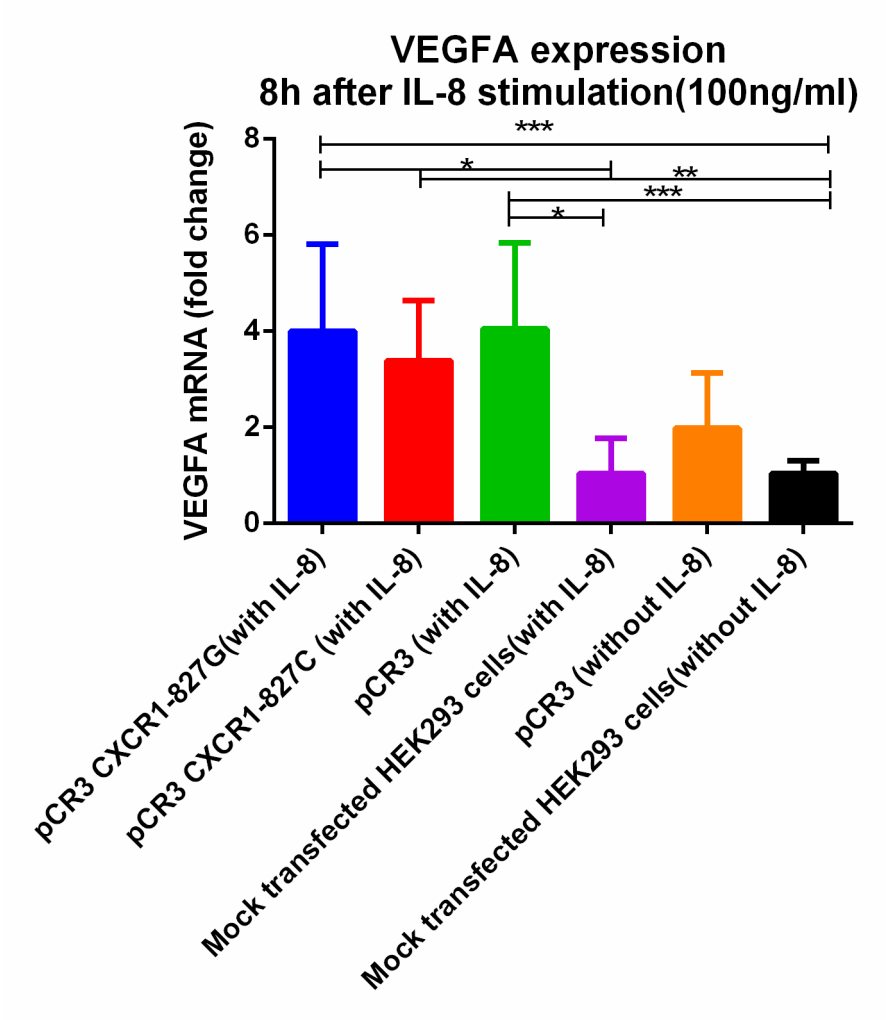


Figure 44 Expression of VEGFA in HEK293 cells transiently transfected with CXCR1 827G/C alleles with and without IL-8 stimulation.

Expression of VEGFA mRNA extracted from IL-8 treated pCR3, pCR3 CXCR1 827G, pCR3 CXCR1 827C transfected and mock-transfected HEK293 cells and non-IL-8 treated pCR3 transfected and mock-transfected HEK293 cells was determined by RT-qPCR. Each condition was repeated at least five times and the qPCR was performed in duplicate for each sample. GAPDH was used as housekeeping controls and the fold change of VEGFA in each test sample compared with IL-8 treated mock transfected cells was calculated by 2-ΔΔCT method as described in **Section 2.3.6.3.1**. Statistical analysis was performed using Kruskal-Wallis test and Dunn's multiple comparisons test. Data are presented as mean ±SD, *P<0.05, **P<0.01, ***P<0.001

5.4.2 Measuring the induction of VEGFA by IL-8 using the stable transduction model

Although induction of VEGFA was observed in the transient transfection models, the effects were likely due to confounding factors instead of the overexpression of CXCR1. In order to make clear the influence of the SNP on VEGFA expression in response to IL-8, the same assay was performed in stably transduced cells to avoid the effects caused by the transfection procedure. 100ng/ml IL-8 was used for the assay and the 8h time point was chosen for the measurement.

As shown in Figure 45 and 46, neither the transduced SiHa cells nor the transduced CaSki cells could significantly increase the expression of VEGFA in comparison with the corresponding parental cells and there was no significant difference observed between CXCR1 827G and CXCR1 827C transduced cells. The results suggested that exogenous IL-8 is unlikely to stimulate VEGFA expression in SiHa and CaSki cells and the SNP does not play a role. Considering that no positive findings were observed in either transiently or stably transduced cells, this was not explored further.

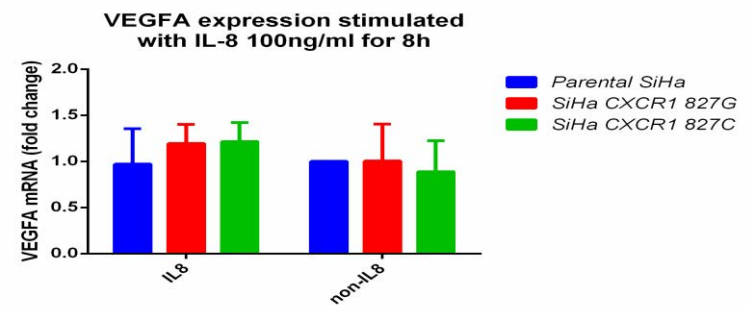


Figure 45 Expression of VEGFA in IL-8 treated SiHa cells transduced with CXCR1 827G/C alleles

CXCR1 827G or CXCR1 827C transduced SiHa cells and parental SiHa cells were incubated with or without 100ng/ml IL-8 for 8h.Afterwards, expression of VEGFA mRNA extracted from each sample was determined by RT-qPCR. Six independent assays were performed to obtain the data and in each assay, qPCR was performed in duplicate for each sample. GAPDH was used as housekeeping controls and the fold change of VEGFA in each test sample compared with IL-8 treated mock transfected cells was calculated by 2-ΔΔCT method as described in Section 2.3.6.3.1. Statistical analysis was performed using ordinary two-way ANOVA test and Sidak's multiple comparisons test. Data are presented as mean ±SD.

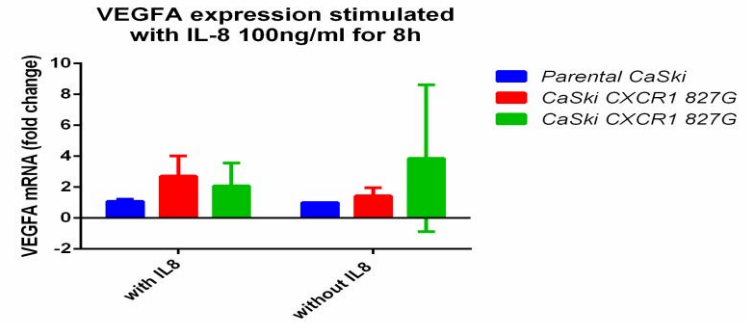


Figure 46 Expression of VEGFA in IL-8 treated CaSki cells transduced with CXCR1 827G/C alleles

CXCR1 827G or CXCR1 827C transduced CaSki cells and parental CaSki cells were incubated with or without 100ng/ml IL-8 for 8h.Afterwards, expression of VEGFA mRNA extracted from each sample was determined by RT-qPCR. Three independent assays were performed to obtain the data and in each assay, qPCR was performed in duplicate for each sample. GAPDH was used as housekeeping controls and the fold change of VEGFA in each test sample compared with IL-8 treated mock transfected cells was calculated by 2-ΔΔCT method as described in Section 2.3.6.3.1. Statistical analysis was performed using ordinary two-way ANOVA test and Sidak's multiple comparisons test. Data are presented as mean ±SD.

5.5 Discussion

Before this study, potential functional roles of CXCR1 rs2234671 have only been described in two studies. In one of the studies, candidemic patients with the minor allele[40] of rs2234671 were shown to be more susceptible to disseminated infections. De-granulation and fungal killing abilities of neutrophils from minor allele carriers were found to be significantly lower than the major allele homozygotes (Swamydas et al., 2016). In the other study, the minor allele was found to be correlated with urine IL-8 levels in women with asymptomatic bacteriuria (Hawn et al., 2009). Apart from these two studies, the roles of CXCR1 rs2234671 in other biological processes and in other disease conditions have not been investigated yet. Therefore, this project focused on the impact of CXCR1 rs2234671 on IL-8 induced cell motility, cell proliferation and VEGFA expression in HR-HPV-related cervical disease.

In the chemotaxis assay, HEK293 cells expressing the minor allele of CXCR1 rs2234671 exhibited significantly higher chemotactic activity towards IL-8 than those cells expressing the major allele. This finding suggested that CXCR1 rs2234671 might be functionally important in the pathogenesis of HR-HPV-related cervical disease by altering the IL-8 induced chemotactic activity of CXCR1-expressing cells. The HEK293 cell line was used in this assay because it was easy-to-transfect and had been shown to work in chemotaxis assays under different experimental conditions (Vinet et al., 2013; Wang et al., 2016b), including IL-8 induced chemotaxis (Boisvert et al., 2012; Feniger-Barish et al., 2003). However, it is noteworthy that this cell line does not express CXCR1 so it might not

40 In this study, subjects with a threonine at the amino acid position 276 of the CXCR1 protein were considered to have CG+GG genotype. The genotypes were called according to the non-coding strand, which did not comply with the conventional nomenclature. The equivalent genotypes in accordance with the conventional nomenclature system should be CG+CC and the minor allele should be C.

be the best cell line to reflect the real IL-8-CXCR1 signalling in physiological and pathological conditions. Thus, the difference between the two alleles of CXCR1 rs2234671 in IL-8 induced chemotaxis has to be discussed cautiously. However, the fact that cell migration could be induced in HEK293 cells transiently transfected with CXCR1 827G/C alleles hints at a real functional difference between the two alleles.

Unlike chronic liver diseases, inflammatory gastrointestinal diseases and glomerular diseases, the distribution of CXCR1 in different stages of HR-HPV-related cervical disease has not been well characterized yet. Since previous studies showed increased CXCR1 expression in cervical cancer tissues, cervical cancer cell lines were chosen to further evaluate the effects of CXCR1 rs2234671 on cell motility in response to IL-8 using the wound healing assay. Penetration of tumour cells through the basement membrane is the hallmark event in cervical cancer development and tumour cell migration is also an essential step in the metastatic process. Similar to the chemotaxis assay, in the wound healing assay, I observed that CaSki cells overexpressing CXCR1 827C have a significantly higher migration activity towards exogenous IL-8 than those expressing the major allele G in both the normal and the serum nutrient deficient culture conditions. These results indicated that the serine to threonine substitution of CXCR1 rs2234671 may functionally contribute to the development and the metastasis of cervical cancer through the enhancement of tumour cell migration in response to IL-8. In addition, it is notable that exogenous IL-8 does significantly enhance the migration capacity of CXCR1 827C transduced cells, but the migration activity of CXCR1 827G transduced cells did not differ before and after exogenous IL-8 stimulation. This finding also provides a hypothetical explanation for the shorter overall survival in cervical cancer patients

with high IL-8 expression, but clearly need to be further investigated. Furthermore, the reason for the absence of response to exogenous IL-8 in CXCR1 827G transduced CaSki cell is currently unknown and requires further study.

Cell proliferation is another concern in the development of cancer. As mentioned previously, IL-8 could stimulate tumour cell proliferation in cervical cancer through its receptor CXCR1. Therefore, proliferation assay were also performed to compare the effects of the two alleles on CXCR1 mediated tumour cell proliferation. In the proliferation assay with CaSki cells, the performance of the CXCR1 827G transduced cells was similar to that of the untransduced parental CaSki cells. Exogenous IL-8 could not increase cell proliferation in either group, compared with the corresponding controls. By contrast, proliferation of CXCR1 827C transduced CaSki cells was significantly increased by exogenous IL-8 stimulation over a very wide range of concentrations (i.e. from 1 to 100ng/ml), compared with untreated controls. Similar to the findings in the wound healing assay, the underlying mechanism for the lack of response to exogenous IL-8 in CXCR1 827G CaSki cell is not clear. But the results of the proliferation assay strongly confirmed the functional difference between the two alleles in the proliferative response towards exogenous IL-8.

The proliferation assay was also performed in SiHa cells. In this assay, the proliferation rate of parental SiHa cells was not significantly enhanced by exogenous IL-8. This finding is slightly different from a previous study performed by Liu and colleagues, in which all concentrations of exogenous IL8, from 1 to 100ng/ml, were shown to enhance SiHa cell proliferation compared with untreated controls, although no significant difference were observed among these three IL-8 treated groups (Liu

et al., 2014). The inconsistency between the two studies might be due to differences in cell culture conditions, sources of the cells and the measurement methods for cell proliferation. Similar to CXCR1 827G transduced CaSki cells, the performance of CXCR1 827G transduced SiHa cells did not differ from parental SiHa cells. Only CXCR1 827C transduced SiHa cells exhibited an enhanced proliferative response towards exogenous IL-8 but only at the concentration of 10ng/ml. Unlike CaSki cells, when treated with 100ng/ml IL-8, proliferation rates of CXCR1 827G transduced and parental SiHa cell groups significantly decreased in comparison with their corresponding 10ng/ml IL-8 treated counterparts. These results suggested that SiHa cells might be less "tolerant" to high concentration of exogenous IL-8 and the decrease of the proliferation rates might be due to IL-8 toxicity or to increased desensitization and internalization of the receptors. However, even in this condition, CXCR1 827C transduced SiHa cells showed better tolerance to IL-8 than the two groups. Again these findings confirmed the functional difference between the two alleles in the proliferative response towards exogenous IL-8.

Due to the differences in genetic background, it is not appropriate to compare SiHa cells and CaSki directly. But in general, SiHa cells exhibited a weaker response towards exogenous IL-8 and the response window seemed narrower than CaSki cells. The intrinsic genotype of SiHa cells at the SNP locus of interest is G/G while CaSki cells have the C/C genotype. It is not clear yet whether the intrinsic genotype could be linked to difference in response between the cell lines, but it is clear that in both assays, only cells overexpressing the minor allele exhibited an increased proliferative response towards exogenous IL-8. Considering that IL-8 is upregulated in high grade cervical intraepithelial lesions and cancers, the difference between the two alleles in

proliferative response towards exogenous IL-8 might be an explanation for the previously observed minor allele enrichment in HPV16 positive cervical cancers.

Furthermore, I also tested the hypothesis that CXCR1 rs2234671 has a functional effect on the induction of VEGFA stimulated by exogenous IL-8 in cervical cancer cell lines but no effect was found. However, since IL-8 has been extensively shown to promote angiogenesis through a direct interaction with endothelial cells, whether the two alleles of the SNP act differently in IL-8 induced angiogenic responses of endothelial cells deserves further study.

Although rs2234671 has been shown to be functional in this project, I could not rule out the possibility that other functional SNPs within the same LD bin also contributed to the association signal observed in the association study. It should also be considered that that SNPs within the same LD bin may have opposite functional effects (Bai et al., 2004; Zeng et al., 2017). It is likely that the observed association and the observed functional effects are the net effect of multiple functional loci in high LD. Thus there is still a need to further functionally characterise other possible disease related SNPs in LD with rs2234671.

Since individuals with the same genotypes may bear different haplotypes, it is also important to pay attention to the functional effects of different combinations of haplotypes. While making the plasmids for the model cellular systems, I noticed that the haplotype carrying all minor alleles of the three markers (i.e. rs16858811, rs2234671 and rs16858808) does not exist in the population. This haplotype may have been eliminated by selection during evolution. Although this missing haplotype has not been investigated in this project, it is a promising target for future functional

assays. We might be able to draw direct inferences from the phenotypic changes led by this haplotype, similar to the use of whole gene knockin or knockout models.

Chapter 6 Discussion

6.1 Summary of findings

It is known since a long time that SNPs can be associated with common diseases. Enormous efforts and money have been invested in hunting for novel susceptible loci and the subsequent functional follow-up studies. However, for HR-HPV-related cervical cancer, findings from previous genetic association studies only explain a small proportion of the heritability. Besides this, very few of the reported associative findings have been further functionally validated or dissected so as to gain novel insights into the mechanisms by which they may contribute to disease. Although updated GWA studies with increased sample sizes and imputation data were predicted to yield new associations, these expectations have not been fulfilled yet. This project was initiated during the wait for new discoveries from genomic approach based studies. In this study, a hypothesis-driven candidate gene approach was used to identify novel functional SNPs associated with HR-HPV-related diseases, and one of the identified SNPs was subsequently investigated with regard to its functional mechanism. The overarching hypothesis was "there are functional SNPs within genes involved in host defence and/or oncogenesis that are associated with HR-HPV-related cervical cancer and precancerous cervical disease". In order to increase the probability to detect genuine associations, candidate genes and SNPs were rationally selected using literature mining methods. The key findings of the project are summarised below

- The SULF1 SNP rs2623047 revealed a strong significant association with the susceptibility to HR-HPV-related cervical disease in the genetic association study, of which the minor allele G appears to be a protective allele against HR-HPV infection.

- The CXCR1 SNP rs2234671 was significantly associated with the development of cervical and vulval cancer in the presence of HPV16. The minor allele C was found to be enriched only in HPV16 positive cervical and vulval cancers in the genetic association study.
- The chemotactic response towards exogenous IL-8 was significantly higher in HEK293 cells overexpressing CXCR1 with the minor allele C of the CXCR1 SNP rs2234671 than those with the major allele G.
- The minor allele C of the CXCR1 SNP rs2234671, rather than the major allele G, resulted in significantly enhanced cell migration of stably transduced CaSki cells in response to exogenous IL-8.
- The minor allele C of the CXCR1 SNP rs2234671, instead of the major allele G, led to significantly increased proliferative responses towards certain concentrations of exogenous IL-8 in stably transduced CaSki cells and SiHa cells.
- Exogenous IL-8 could not induce VEGFA expression in SiHa and CaSki cells and the SNP CXCR1 rs2234671 did not affect this.

6.2 Discussion and future directions

6.2.1 Candidate gene approach in HR-HPV-related cervical disease

In recent years, high rates of spurious associations and non-replication of previous findings markedly dampened researchers' enthusiasm for pursuing functional disease related SNPs through a candidate gene approach. Since candidate gene based association studies could be conducted with ease, in numerous studies, casually selected genes and SNPs have frequently been tested. However, in such situations,

even if association was identified, it was likely to be caused by chance or by linkage with unknown causal SNP. Findings obtained in this way are usually considered less convincing than those of GWA studies. Even worse, functional interpretation and translation of these findings is no easier than those from GWA findings. This directly leads to the general impression that the candidate gene approach is inferior to genomic approaches and consequently hinders the application and the improvement of candidate gene based genetic association studies. Nevertheless, the study presented in this thesis provides evidence that a rational candidate gene approach could still be a powerful and efficient way to identify novel disease related SNPs and to proceed with functional translations, especially when an appropriate candidate gene and SNP selection strategy was applied to well characterised clinical samples from different stages of disease.

The majority of susceptible loci previously identified in HR-HPV-related cervical disease are located in immune genes and genes known to be involved in carcinogenesis, which indicates that genetically-mediated alterations in host defence responses may play an essential role in the susceptibility to HR-HPV infection and cervical cancer (Mehta et al., 2017). In this project, the identification of the two SNPs associated with HR-HPV-related cervical disease provides further evidence for this indication. The associations between the two SNPs and HR-HPV related cervical disease still need to be validated in a larger sample set. As exploration deepens, more immune response gene polymorphisms are expected to be discovered and linked with HR-HPV-related cervical disease.

As mentioned previously, since cervical cancer is a multistep disease, different SNPs may be involved in different transition states. The design of the case-control study,

especially the grouping strategy, directly determines what results could be achieved. In this project, the two key SNPs identified were found to be associated with different transition states of HR-HPV-related cervical disease. SULF1 rs2623047 was associated with the acquisition of HR-HPV while CXCR1 rs2234671 was associated with the progression from HSIL to invasive cancer in the presence of HPV16. However, it is noteworthy that both associations were identified in the validation step, rather than the original discovery phase.

Due to the poor performance of genomic DNA extracted from formalin-fixed, paraffin-embedded cancer samples, the original array could only compare the frequency of alleles or genotypes between HR-HPV negative, cytologically normal controls and HR-HPV positive, HSIL cases. This design was surely not ideal since it was not able to identify any SNPs that are involved in the transition from HSIL to cancer. I have to admit that the association between CXCR1 rs2234671 and the progression of the disease was picked out quite by chance. It would have been missed in the discovery phase if the genotyping error did not occur. The findings for CXCR1 rs2234671 suggest the possibility that the same array panel may yield new associations when re-evaluated in case-control studies including the transition from HSIL to cancer. By contrast, the association signal for SULF1 rs2623047 was successfully captured by the original array during the discovery phase. However, it left unanswered in this phase the question in which transition state the SNP got involved. As mentioned above, at least two noticeable transition states exist during the development of HSIL. Not until the additional samples had been genotyped did I find out that SULF1 rs2623047 was associated with HR-HPV infection rather than the formation of HSIL.

Based on the findings in this project, each candidate SNP needs to be separately evaluated in each transition state so as to draw a more precise conclusion. An ideal case-control candidate gene based study in HR-HPV-related cervical disease should start with a comparison between HR-HPV negative, cytologically normal controls and HR-HPV positive cancer cases. This kind of design ensures that all the known transition states are included and increases the power to detect association in the discovery phase. In the following step, HR-HPV positive, cytologically normal or LSIL cases and HR-HPV positive, HSIL cases should be genotyped to further delineate associations among each transition states.

6.2.2 A potential follow-up study for SULF1 SNP rs2623047

SULF1 rs2623047 has been found to be associated with HR-HPV infection in this project and the association is independent from previously reported association signals within intronic regions of SULF1 (Wang et al., 2010). Since functional assays for this SNP could not be done within the timeframe of this project, the *in silico* tool HaploReg v4 (Ward and Kellis, 2011) was used alternatively to generate mechanistic hypotheses for future functional analysis. In total, 8 other SNPs have been found in strong LD with rs2623047 (see Figure 47). Of the 8 SNPs, two are located upstream of the SULF1 gene while the rest all reside in intronic regions. The two upstream SNPs, rs2704027 and rs2244817, are in complete LD with rs2623047.

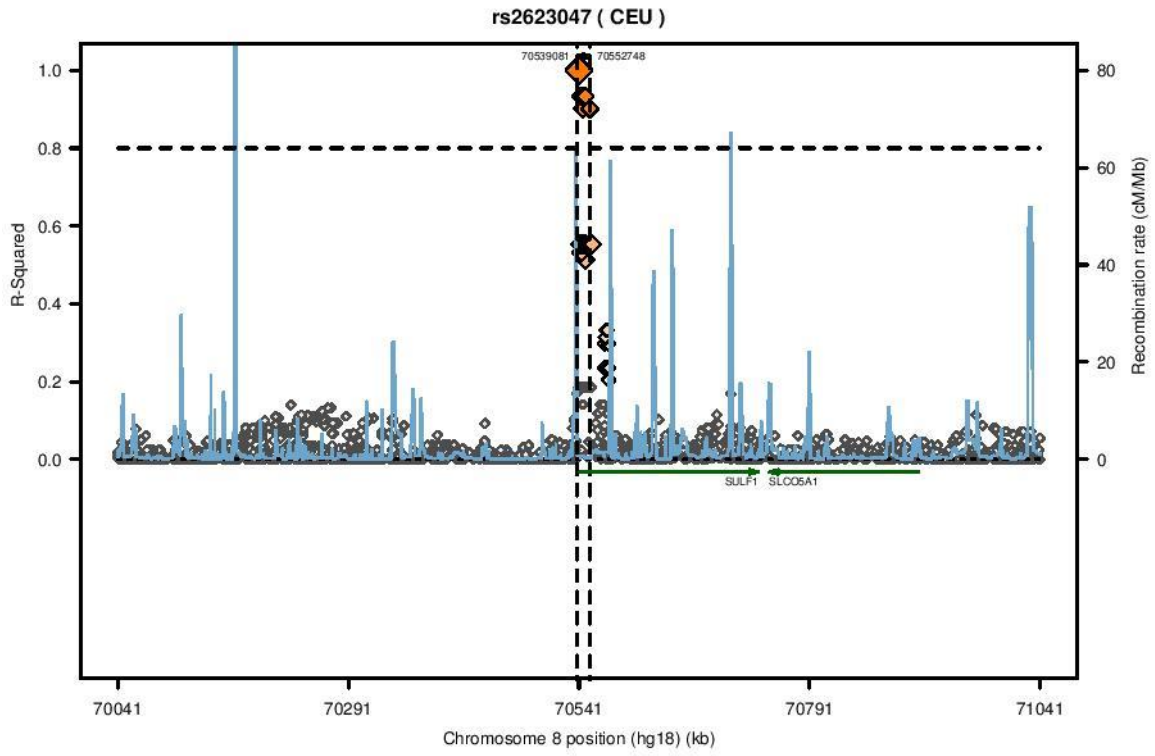


Figure 47 Regional LD plot of SULF1 rs2623047

This plot was generated using online tool SNAP http://archive.broadinstitute.org/mpg/snap/ldplot.php based on 1000 Genomes pilot 1 database (CEU).The lead SNP (rs2623047) is represented by the largest diamond in the middle of the plot. The 8 SNPs in strong LD with the lead SNP ($r^2 \geq 0.8$) are represented by smaller diamonds above the horizontal black dotted line, including rs2704027, rs2244817, rs1441201, rs1899276, rs2725090, rs2725089, rs2005370 and rs1545219.

According to the chromatin signals and transcription factor binding profiles provided by the HaploReg v4 database, the lead SNP rs2623047 should be prioritized as the top candidate SNP in the LD bin for further functional follow-up studies. This SNP overlaps with ChromHMM-predicted promoter histone marks in 12 major tissue types, including cervical tissue. In the HeLa-S3 cervical carcinoma cell line, the rs2623047 locus is classified as an active transcription start site by the 15-state core chromatin state model and as a weak enhancer by the 25-state model with imputed marks. This SNP has been previously associated with differential SULF1 expression

in cerebellum and temporal cortex tissues by a eQTL study (Zou et al., 2012). The HaploReg v4 tool also predicted that rs2623047 could change the match to transcription factors (TFs) TCF12_disc6 and TEF-1_2. The minor allele G has a higher predicted relative affinity for both TFs mentioned than the major allele A and the change in log-odds (LOD) score for TCF12_disc6 is large. Based on these lines of evidence and the previous study showing higher promoter activity of the G allele than the A allele in an ovarian cancer cell line (Han et al., 2011), it is reasonable to speculate that the SNP could affect susceptibility to HR-HPV infection functionally through regulatory mechanisms. The G allele induced changes in TF binding affinity may lead to increased SULF1 expression, altered modulation of HSPGs on cervical epithelial cells by sulfatase 1, and finally reduced attachment of HR-HPV to HSPGs, which might affect the ability of the virus to enter epithelial cells. However, this is just a hypothetical explanation for the observed association in this project. Whether the SNP affects the binding affinity of above mentioned TFs in cervical epithelial cells, whether SULF1 is differentially expressed in cervical tissue of G allele carriers, whether and how the above mentioned TFs affect the expression of SULF1 or other functions, need to be answered in future functional assays. SNP rs2244817 should also be included in such functional assays since the locus has been associated with CTCF and RAD21 protein binding in ChIP-Seq experiments.

6.2.3 A functional disease related SNP CXCR1 rs2234671

The association of CXCR1 rs2234671 with disease progression is the key finding of this project and will be the main focus for the future work. CXCR1 rs2234671 was found to be associated with the progression from high-grade intraepithelial lesion to

invasive cancer in the presence of HPV16. The minor allele of this SNP was found to be responsible for enhanced cell migration and proliferation in HPV16 positive cervical cancer cells in response to exogenous IL-8. These findings provide possible explanations for the observed associations, but clearly need to be further investigated.

Firstly, the question "why was the enrichment of the minor allele of the SNP only observed in HPV16 positive cervical and vulval cancers, but not in other HR-HPV positive cancers" remains unanswered in this project. A possible explanation is the interaction between HPV16 and the ligand IL-8. Both HPV16 E6 and E7 have been shown to increase IL-8 expression in primary foreskin keratinocytes (Walker et al., 2011). Abnormal cervical or vulval epithelial cells of the minor allele carriers may have higher migratory and proliferative responses to IL-8 induced by HPV16 and therefore be more likely to break through the basal membrane and invade adjacent organs or tissues. Although increased IL-8 expression in HSIL and cervical cancer has been shown in published studies and the previous study in our group, it is still unclear whether this increase is HPV type specific. If the increase of IL-8 is universal, this hypothesis will be rejected. Another possible explanation is the direct interaction between HPV16 and CXCR1. In a previous genetic association study, polymorphisms within FGFR2, HSPG2 and SDC2 were found to be associated with HPV16 induced cervical lesion progression. All these genes had been shown to directly interact with HPV16 (Zou et al., 2016). Similarly, CXCR1 may directly interact with HPV16 in a so far unelucidated carcinogenic process. The interaction between another chemokine receptor CXCR4 and HPV16 viral proteins has been reported (Cheng et al., 2014). However, no study has ever investigated the direct

interaction between HPV16 and CXCR1. Thus, CXCR1 and HPV16 interaction could also be a focus for further investigations. In this project, stable transduction and subsequent functional assays were performed in HPV16 positive cervical cancer epithelial cells. It would be interesting to study the functional effects of the SNP in HPV16 negative cervical epithelial cells or in HPV18 positive cells in future studies since this may give a hint on the processes in which the interaction participates. Additionally, the possibility can not be ruled out that the interaction between the SNP and other HR-HPV types was missed because of the inadequate number of other HR-HPV positive cancer cases in this study. Whether the SNP associate with other HR-HPV type related cervical disease still needs to be verified in larger sample set.

Secondly, the functional effects of the SNP identified in this project add to the evidence that IL-8-CXCR1 signalling plays a critical role in the development of HR-HPV-related cervical and vulval cancers. Since IL-8 and its receptors have been shown to be multifunctional in previous studies on other diseases, undoubtedly they will participate in a wide range of biological processes in HR-HPV-related disease. However, at the moment, our knowledge about this signalling in HR-HPV-related diseases remains limited. It is still not well elucidated how IL-8 and CXCR1 distribute in different cells and tissues and what processes they get involved in during the diseases courses. And according to the results presented here, it is unclear whether exogenous IL-8 and endogenous IL-8 have the same effects on target tissues and cells. To better understand the functional effect of the CXCR1 rs2234671 SNP, especially in different cell types and tissues, a better background knowledge at the genetic level is necessarily required.

Apart from the results obtained from the functional assays, there was another finding that aroused my interest in gene functions. As mentioned previously, an evolutionary survey study considered that the serine to threonine change at amino acid position 276 of CXCR1 was the consequence of accelerated adaptive evolution unique to the human lineage (Liu et al., 2005). However, after comparing the nucleotide sequence of CXCR1 and CXCR2, I think there might be another explanation.

E3 loop amino acid sequence
CXCR1:DTLMRTQVIQE**S**CERRNNIGR
CXCR2:DTLMRTQVIQE**T**CERRNHIDR

E3 loop nucleotide sequence
CXCR1:gacaccctcatgaggacccaggtgatccaggag**agc**tgtgagcgccgcaacaacatcggccgg
CXCR2:gacaccctcatgaggacccaggtgatccaggag**acc**tgtgagcgccgcaatcacatcgaccgg

Figure 48 Nucleotide and amino acid sequence alignments of the E3 loop in CXCR1 and CXCR2

This figure shows the nucleotide sequence and amino acid sequence of the third extracellular loop for both CXCR1 and CXCR2. The rs2234671 locus on CXCR1 and its homologous locus on CXCR2 are highlighted in bold and underlined.

As can be seen from Figure 48, the sequences encoding the third extracellular loop of CXCR1 and CXCR2 reveal 93.7% nucleotide identity. Although an accession ID rs201881186 has been issued to the homologous locus of CXCR1 rs2234671 within CXCR2 by dbSNP, it is not considered to be an actual SNP since all subjects (except for one Indian subject) in the 1000 Genomes project (phase 3) have the C/C genotype at this locus. By contrast, only 3.2% of these subjects have a C/C genotype at the CXCR1 rs2234671 locus. Considering that the C allele of rs2234671 has been consistently reported as the risk allele in different disease conditions, it is reasonable to speculate that the C allele of rs2234671 is the actual ancestral allele, just like the

rs201881186 locus, and that the allele frequency dropped drastically during the evolution as a result of negative selection, rather than the other way round as mentioned in the published study. If this speculation is true, it will add to the evidence that the two alleles are functionally different. Besides, it may indicate that the deleterious effects of the minor allele C of rs2234671 will be more pronounced or only exist in processes that mediate solely by CXCR1, rather than CXCR2. In order to further investigate the functional differences of the two rs2234671 alleles, functions exclusively mediated by CXCR1 need to be investigated. Currently, however, functional differences between the two IL-8 receptors in HR-HPV-related diseases have not been fully understood. The functional roles of CXCR1 and CXCR2 need to be dissected first to narrow down the list of functions that can be attributed to the CXCR1 rs2234671 SNP association with cervical cancer.

In addition, although this project provides evidence that SNP rs2234671 can affect exogenous IL-8 induced cell migration and proliferation, the underlying mechanisms of these phenotypic changes remain unknown. It is currently not known whether rs2234671 could alter ligand binding activity or cause a conformational change of the encoded receptor (or both). Considering that the encoded amino acid of this SNP locates next to a previously mentioned disulfide bond Cys30-Cys277 in CXCR1, together with the evidence showing the importance of the third extracellular loop of CXCR1 in the binding interaction (Barter and Stone, 2012), the role of the SNP in IL-8-CXCR1 interaction might well repay detailed investigation.

Ideally, in follow-up studies, investigations on gene functions and the SNP functions could be performed in parallel. As the knowledge about CXCR1 extends, more novel insight into the functions of the rs2234671 could be gained, and vice versa. Clearly

numerous questions about the rs2234671 and CXCR1 have yet to be answered. Some of these questions might be dealt with in near future, such as whether endogenous IL-8 causes the baseline change in current wound healing assay, whether the SNP affects IL-8 induced angiogenesis in endothelial cells, and what functional impacts the missing haplotype (i.e. formed by the minor alleles of rs16858811, rs2234671 and rs16858808) have.

Overall, this project has demonstrated the feasibility of using a candidate gene approach to identify potentially functional SNPs in HR-HPV-related diseases. It has also refined the knowledge of which samples should be included in such a study at both discovery and validation phases. Two novel SNPs (SULF1 rs2623047 and CXCR1 rs2234671) have been found to be associated with different transition states in the pathogenesis of HR-HPV related cervical diseases and both are promising targets for further functional follow-up studies. Results from the association study and functional assays suggest that CXCR1 rs2234671 may affect cervical cancer susceptibility by altering the function of the IL-8-CXCR1 signalling in the presence of HPV16. If the genetic association could be further validated in a larger sample set, in conjunction with screening programmes, this SNP may be used to improve triage of "at risk" patients with HR-HPV infection. As HR-HPV infection is becoming increasingly associated with other forms of cancer including penile, anal and head and neck cancers, the findings may also be applicable more generally. In addition, since CXCR1 rs2234671 has been shown to be associated with many other diseases (including, but not exclusively other cancers), the functional effects identified in this project for this SNP may also be used to explain the pathogenesis of these other diseases.

Appendix

Appendix-1 Sample size estimation for the SNP array under the log-additive model

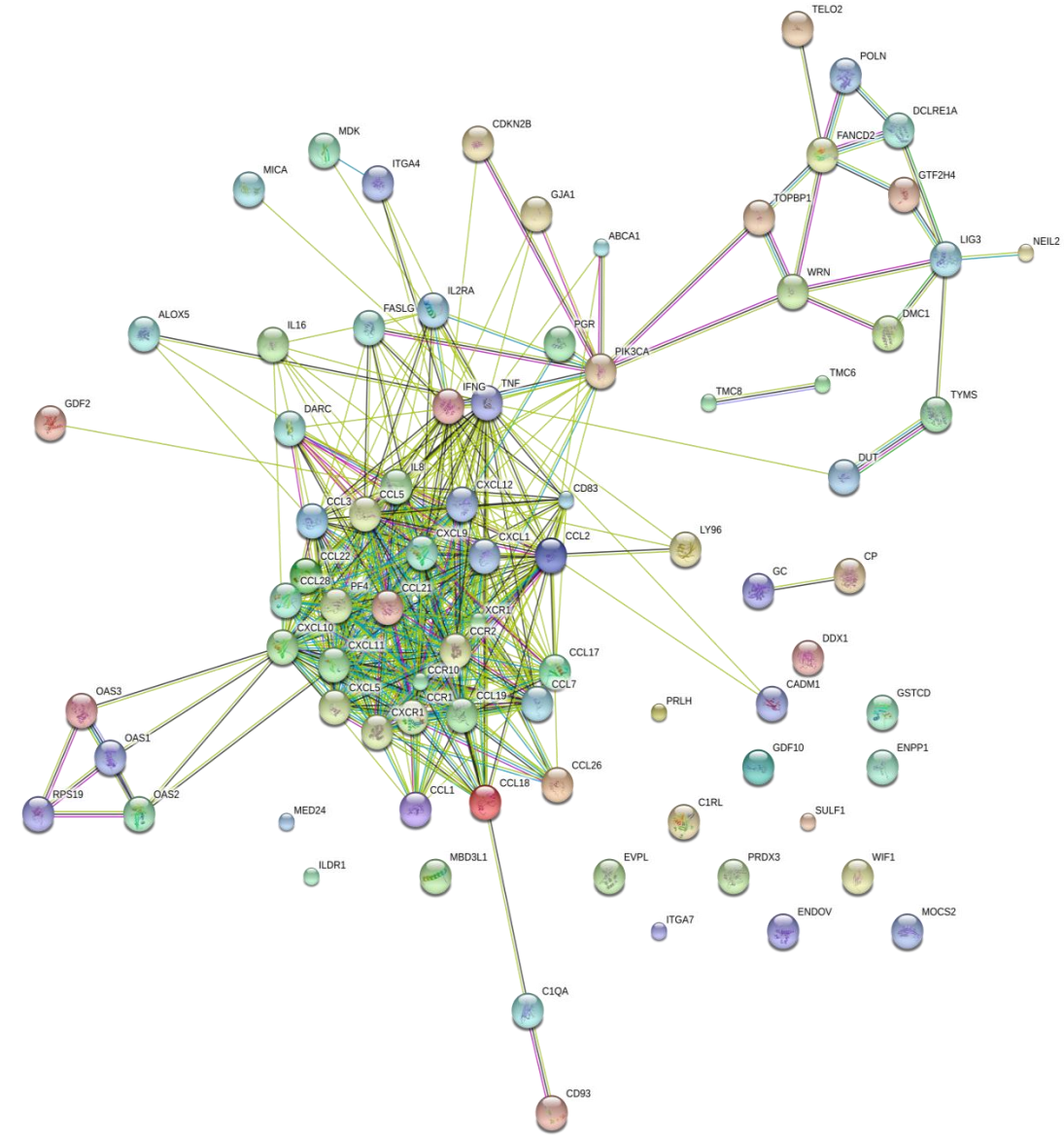
Outcome:	*Disease*
Design:	*Matched case-control*
Hypothesis:	*Gene only*
Desired power:	*0.800000*
Significance:	*0.050000, 2-sided*
Gene	
Mode of inheritance:	*Log-additive*
Allele frequency:	*0.0500 to 0.4000 by 0.1000*
Disease model	*Summary parameters*
P_0 0.000095	*k_P 0.000100*
R_G: 1.5000	*(*indicates calculated value)*

Parameter	Null	Full	Reduced
Gene	$\beta_G=0$	β_G	——

		N	
Frequency	R_G	Gene	P_0
0.050000	1.5000	843	0.000095
	2.0000	257	0.000091
	2.5000	136	0.000087
	3.0000	89	0.000083
0.150000	1.5000	331	0.000087
	2.0000	106	0.000076
	2.5000	59	0.000067
	3.0000	40	0.000059
0.250000	1.5000	236	0.000079
	2.0000	79	0.000064
	2.5000	45	0.000053
	3.0000	32	0.000044
0.350000	1.5000	203	0.000072
	2.0000	70	0.000055
	2.5000	42	0.000043
	3.0000	30	0.000035

N is the number of case-control pairs required for the desired power

Appendix-2 The STRING network of known protein-protein interactions among the 83 candidate genes



Known Interactions	Predicted Interactions	Others
from curated databases	*gene neighborhood*	*textmining*
experimentally determined	*gene fusions*	*co-expression*
	gene co-occurrence	*protein homology*

Appendix-3 Published SNPs within the 83 candidate genes/regions

gene	SNPs	SNPs location	Allele	MAF	Validation	Disease/phenotype	PMID	published paper #
CCL1	rs2282691	intron variant	A/T	T=0.4205/2105	YES	AECOPD/COPD	18550614 16864710	3
	rs3138031	intron variant	A/C	C=0.0835/418	YES	TB	19057661	1
	rs159271	intron variant	C/T	T=0.0972/486	YES	Viseral leishmaniasis	17122780	1
	rs159273	upstream variant 2KB	G/T	T=0.0789/395	YES	Viseral leishmaniasis	17122780	1
CCL2	rs2857656	upstream variant 2KB	C/G	C=0.4209/2108	YES	TB,Carotid atherosclerosis, Non contact soft tissue injuries	18940815 19506371 23890452	7
	rs2857657	intron variant	C/G	G=0.0950/476	YES	Knee osteoarthritis, risperidone treatment	23211090 24495780	6
	rs4586	synonymous codon	C/T	T=0.4577/2292	YES	AMD, risperidone treatment	23848218 24495780	12
	rs13900	utr variant 3 prime	C/T	T=0.3632/1819	YES	Altered preexercise CK levels	20339010	6
	rs3760396	upstream variant 2KB	C/G	C=0.0940/470	YES	NSCLC,rosiglitazone or pioglitazone treatment	21514686 18996102	3
CCL3	rs8951	intron variant, utr variant 3 prime	C/T	C=0.1829/915	YES	HIV	16773571	1
	rs1719130	intron variant	C/T	C=0.2382/1193	YES	HIV	20725607 16773571	2
	rs1719134	intron variant, nc transcript variant	A/G	A=0.2220/1112	YES	HIV	16773571	3
	rs1130371	intron variant, synonymous codon	C/T	A=0.2210/1106	YES	HIV/HIV associated Dementia	20725607 16773571	3
CCL5	rs2107538	upstream variant 2KB	C/T	T=0.3077/1541	YES	PTC,PTDM	23957698 20805685	23
	rs1065341	utr variant 3 prime	A/G	C=0.1258/630	YES	allergic rhinitis	23663310	1
	rs2280789	intron variant	C/T	G=0.1853/928	YES	PTDM	20805685 23985723	9
CCL7	rs991804	NIL	A/G	T=0.4109/2057	YES	Crohn's disease	20570966	7
CCL17	rs223828	intron variant	C/T	T=0.2308/1156	YES	Classical Hodgkin lymphoma	24008079	1
	rs223899	intron variant	A/C	A=0.0509/254	YES	Prostate cancer	23920401	1
	rs4784805	intron variant	A/C	T=0.2308/1156	YES	Kawasaki disease	23942559	1

CCL18	rs712043	intron variant	C/T	T=0.2033/1018	YES	HIV	16773571	1
	rs854472	intron variant	A/G	C=0.2053/1028	YES	HIV	16773571	1
	rs2015070	intron variant	A/G	T=0.1058/530	YES	HIV	16773571	1
	rs2015086	upstream variant 2KB	C/T	G=0.2548/1275	YES	Higher macrophage expression of CCL18	19036375	1
	rs14304	NIL	A/G	T=0.2043/1022	YES	Breast cancer	21935604	1
CCL19	rs2233872	intron variant, upstream variant 2KB	C/T	G=0.1687/844	YES	MI	24493450	1
CCL21	rs2812377	intron variant, upstream variant 2KB	G/T	C=0.2796/1400	YES	Coronary artery disease	24990231	1
	rs2812378	intron variant, upstream variant 2KB	C/T	G=0.3013/1509	YES	RA,Endometriosis	18794853 20797713	10
	rs11574914	intron variant, upstream variant 2KB	C/T	A=0.2001/1002	YES	MI	24493450	3
	rs11574915	intron variant, utr variant 5 prime	G/T	C=0.1022/511	YES	MI	24493450	1
CCL22	rs170359	intron variant	A/G	G=0.2364/1184	YES	Prenatal IgE	19968631	1
	rs4359426	missense	A/C	A=0.0835/417	YES	Atopic dermatitis	22125604	1
CCL26	rs2240478	intron variant	C/T	A=0.2393/1197	YES	Asthma	17848170	2
	rs2302009	utr variant 3 prime	G/T	C=0.2786/1394	YES	Asthma,RA	17848170 15784470	2
	rs6965556	intron variant	C/T	T=0.2899/1451	YES	RA	15784470	2
CCL28	rs11950448	intron variant	C/G	G=0.3872/1939	YES	Family chaos	18360741	1
CXCL1	rs4074	intron variant	A/G	G=0.4044/2024	YES	HCV infection	22173151	3
	rs117604	upstream variant 2KB	C/T	T=0.4728/2367	YES	Ischemic stroke	23198952	1
CXCL4	-	-	-	-	-	-	-	-
CXCL5	rs352046	upstream variant 2KB	C/G	G=0.1997/1000	YES	DM,ACS	19035625 18769620	4
	rs425535	synonymous codon	A/G	T=0.1869/936	YES	-	-	3

CXCL8	rs4073	upstream variant 2KB	A/T	T=0.4778/2392	YES	CRC,Sepsis,MI,NSCLC,Oral cancer,Apical periodontitis, Idiopathic pulmonary fibrosis, Childhood IgA nephropathy, Oral lichen planus,NHL	25139485 25000179 24462138 23831257 23545310 22788685 21649933 21214373 19842025 19066394	43
	rs1126647	utr variant 3 prime	A/T	T=0.2851/1428	YES	-	-	2
	rs2227306	intron variant	C/T	T=0.2594/1299	YES	NSCLC,RA,Childhood IgA Nephropathy,Oral lichen Planus,GD,asthma	23831257 21385363 21214373 19842025 19816813 16503988	16
	rs2227307	intron variant	G/T	G=0.4235/2120	YES	Periodontitis	21188496	8
	rs2227543	intron variant	C/T	T=0.2885/1444	YES	Childhood IgA nephropathy	21214373	3
CXCL9	rs10336	intron variant, utr variant 3 prime	C/T	A=0.2997/1500	YES	Chagas cardiomyopathy	23150742	4
	rs2276886	intron variant	A/G	T=0.2065/1034	YES	Electrocardiographicvariability	17903306	1
	rs3733236	intron variant, utr variant 3 prime	C/T	A=0.1426/714	YES	T1DM	19410617	1
CXCL10	rs3921	intron variant, utr variant 3 prime	C/G	C=0.3075/1540	YES	Bone marrow transplantation	23291247	3
	rs8878	intron variant, utr variant 3 prime	C/T	A=0.3085/1544	YES	Colorectal cancer,T1DM	24748971 19410610	4
	rs4859587	intron variant	A/C	A=0.3087/1545	YES	-	-	2
	rs4859588	intron variant	A/G	G=0.3087/1545	YES	Sepsis in premature neonate	25000179	1
	rs35795399	intron variant, utr variant 3 prime	C/T	C=0.0066/33	YES	T1DM	19410617	1
CXCL11	rs6817952	intron variant	A/G	A=0.1607/805	YES	Inflammatory bowel disease	20848514	1

CXCL12	rs1804429	utr variant 3 prime	G/T	C=0.0587/294	YES	NSCLC	21514686	1
	rs1065297	utr variant 3 prime	C/T	G=0.0661/330	YES	ITP	23078136	1
	rs266093	utr variant 3 prime	C/G	C=0.3480/1742	YES	Cervical diseases	19788587	3
	rs2839696	utr variant 3 prime	C/T	A=0.0585/293	YES	-	-	1
	rs1029153	utr variant 3 prime	C/T	G=0.2504/1253	YES	-	-	2
	rs169097	utr variant 3 prime	A/C/G/T	A=0.0537/268	YES	-	-	1
	rs1801157	utr variant 3 prime	A/G	T=0.1855/928	YES	Multiple myeloma, ALL, Breastcancer, Esophagogastric cancer,NHL,MI,HIV,Hepatocellular carcinoma, Premature ovarian failure	23711392 23653000 21643956 19927352 19821058 19601773 19327121 21296802	35
	rs266089	intron variant	A/G	A=0.1805/903	YES	AD	16385451	1
	rs197452	intron variant	C/T	T=0.1238/620	YES	-	-	1
	rs266087	downstream variant 500B,intron variant	A/G	A=0.3678/1842	YES	-	-	3
	rs2297630	intron variant	A/G	A=0.1839/921	YES	-	-	4
	rs2839693	intron variant	A/G	T=0.1683/843	YES	-	-	2
	rs2839689	intron variant	A/T	A=0.1156/578	YES	-	-	1
	rs3780891	intron variant	A/G	A=0.0675/337	YES	-	-	1
	rs17885289	upstream variant 2KB	A/G	T=0.1805/903	YES	Cervical diseases	19788587	1
	rs2839695	intron variant, utr variant 3 prime	C/T	G=0.1148/574	YES	-	-	1
	rs2839685	upstream variant 2KB	C/T	A=0.1745/874	YES	-	-	1
	rs266085	intron variant	C/T	T=0.3738/1872	YES	Immune thrombocytopenia, Cervical carcinoma	23078136 19788587	3
IL16	rs859	utr variant 3 prime	A/G	G=0.3730/1867	YES	NHL	19066394	4
	rs11325	utr variant 3 prime	G/T	T=0.2506/1254	YES	-	-	4
	rs1131445	downstream variant 500B,utr variant 3 prime	C/T	C=0.2879/144	YES	Graves' disease, Prostate cancer	18394967 24061634	5

IL16	rs1803275	nc transcript variant,synonymous codon	A/G	A=0.1462/731	YES	Alopecia areata	24320753	1
	rs3848180	intron variant	G/T	G=0.4906/2456	YES	CAD	23195133	2
	rs4072111	missense, nc transcript variant	A/G	T=0.1645/824	YES	SLE,CRC	19298795 19073878	6
	rs4128767	intron variant	C/T	G=0.4347/2176	YES	GD	24694201	1
	rs4577037	intron variant	G/T	G=0.1354/677	YES	CAD	23195133	1
	rs4778641	utr variant 3 prime	C/T	C=0.4780/2393	YES	GD	18394967	1
	rs4778889	intron variant, upstream variant 2KB	C/T	C=0.2508/1256	YES	Endometriosis,SLE,GD	20662556 19298795 18394967	10
	rs7175701	intron variant	C/T	C=0.2496/1250	YES	Prostate cancer	22923925	1
	rs7182786	upstream variant 2KB	A/G	G=0.4613/2310	YES	GD	24694201	1
	rs8028364	intron variant	C/G	G=0.4381/2193	YES	GD	24694201	1
	rs8031107	nc transcript variant,synonymous codon	A/G	G=0.4145/2076	YES	GD	24694201	1
	rs8034928	intron variant	C/T	C=0.2224/1113	YES	CAD	23881440	2
	rs11073001	nc transcript variant, synonymous codon	A/G	G=0.2808/1406	YES	Alopecia areata	24320753	1
	rs11556218	missense, nc transcript variant	G/T	G=0.1464/733	YES	Nasopharyngeal carcinoma,CAD,Prostate cancer,SLE,CRC	24101193 23881440 22923025 19298795 19073878	12
	rs11638444	intron variant	C/T	C=0.1873/937	YES	-	-	1
	rs12437819	intron variant	G/T	T=0.2302/1153	YES	-	-	1
	rs12438640	intron variant, upstream variant 2KB	A/G	A=0.2362/1182	YES	-	-	1
	rs12907134	intron variant	A/G	A=0.4343/2174	YES	GD	24694201	1
	rs17875486	intronvariant, upstream variant 2KB	C/T	T=0.3285/1644	YES	Alopecia areata	24320753	1
	rs17875491	intron variant, upstream variant 2KB		C=0.1719/861	YES	Alopecia areata	24320753	1

Midkine	-	-	-	-	-	-	-	-
ALOX5	rs702366	intron variant	A/G	G=0.4908/2458	YES	late-onset Alzheimer disease	16385451	1
	rs745986	intron variant	A/G	G=0.1873/937	YES	subclinical atherosclerosis	20592751	1
	rs892690	intron variant	A/G	T=0.3077/1541	YES	asthma	19214143	1
	rs892691	intron variant	A/G	A=0.2871/1438	YES	atherosclerosis	19130089	4
	rs934187	intron variant	C/T	G=0.2596/1300	YES	late-onset Alzheimer disease	16385451	1
	rs1059696	intron variant	A/G	A=0.1510/755	YES	subclinical atherosclerosis	20592751	1
	rs1369214	intron variant	A/G	G=0.3992/1998	YES	atherosclerosis	19130089	3
	rs1487562	intron variant	C/T	T=0.1661/831	YES	atherosclerosis	19130089	2
	rs1565096	intron variant	A/G	G=0.2083/1043	YES	rhinosinusitis	18366797	2
	rs1864414	intron variant	A/C	G=0.1793/897	YES	Ovarian cancer	22282663	2
	rs2029253	intron variant	A/G	G=0.3960/1982	YES	ACS,atherosclerosis,ischemic stroke	18398223 19130089 23079278	5
	rs2115819	intron variant	C/T	G=0.4816/2411	YES	TB,asthma,subclinical atherosclerosis,atherosclerosis	23448388 20592916 20592751	12
	rs2228064	synonymous codon	A/G	A=0.1212/606	YES	-	18248681	1
	rs2229136	downstream variant 500B, intron variant, synonymous codon	A/G	G=0.0867/433	YES	-	-	1
	rs2242332	intron variant	C/T	T=0.3275/1639	YES	atherosclerosis	19130089	1
	rs2242334	intron variant	G/T	G=0.3271/1638	YES	atherosclerosis	19130089	2
	rs2279435	intron variant	A/C	C=0.3904/1954	YES	rhinosinusitis	18366797	1
	rs2291427	intron variant	A/G	A=0.2909/1457	YES	glioblastoma,Glioma	24005813 19423540	9
	rs3740107	intron variant	A/G	A=0.2780/1391	YES	atherosclerosis	19130089	2
	rs3780894	intron variant	A/G	G=0.1751/877	YES	rhinosinusitis	18366797	1
	rs3780901	intron variant	C/T	C=0.3612/1808	YES	subclinical atherosclerosis	20592751	3
	rs3780902	intron variant	A/G	G=0.0819/409	YES	atherosclerosis	19130089	1
	rs3780906	intron variant	A/G	A=0.3359/1682	YES	subclinical atherosclerosis	20592751	2
	rs3824612	intron variant	C/T	T=0.3079/1541	YES	-	-	2
	rs3824613	intron variant	C/T	T=0.1757/880	YES	atherosclerosis	19130089	1
	rs4540799	intron variant	C/T	T=0.2590/1297	YES	bone mineral density	17909879	1

ALOX5	rs4948672	intron variant	C/G	C=0.4217/2111	YES	rhinosinusitis	18366797	1
	rs4986832	upstream variant 2KB	A/G	A=0.1791/896	YES	montelukast treatment	17460547	2
	rs4987105	synonymous codon	C/T	T=0.1709/856	YES	montelukast treatment	17460548	3
	rs6593482	upstream variant 2KB	G/T	T=0.1791/896	YES	breast cancer	18843019	1
	rs7099684	intron variant	A/T	A=0.1833/918	YES	-	-	3
	rs7894352	intron variant	A/G	A=0.1865/934	YES	bone mineral density	17909879	1
	rs7913948	upstream variant 2KB	A/G	A=0.1791/896	YES	MI	19787205	1
	rs7919239	intron variant	A/C	A=0.2246/1125	YES	rhinosinusitis	18366797	1
	rs10900213	intron variant	G/T	T=0.4964/2486	YES	bone mineral density	17909879	1
	rs10900215	intron variant	C/G	C=0.1322/661	YES	atherosclerosis	19130089	2
	rs11239516	intron variant	G/T	T=0.1010/505	YES	bone mineral density	17909879	1
	rs11239523	intron variant	C/T	C=0.0857/428	YES	rhinosinusitis	18366797	1
	rs11239524	intron variant	G/T	G=0.0964/483	YES	subclinical atherosclerosis	20592751	1
	rs12264801	intron variant	A/G	A=0.4425/2215	YES	rhinosinusitis	18366797	1
	rs12762303	upstream variant 2KB	C/T	C=0.1791/896	YES	CAD,MS	18369664 18366677	2
	rs7090328	intron variant	A/G	A=0.2131/1066	YES	-	-	1
	rs3780905	intron variant	C/T	T=0.1360/680	YES	-	-	1
ABCA1	rs363717	utr variant 3 prime	A/G	C=0.1424/712	YES	TrPN	21245421	3
	rs1799777	intron variant, utr variant 5 prime	/G	C=0.1418/709	YES	HDLC level	19041386	1
	rs1800976	upstream variant 2KB	C/G	G=0.4940/2474	YES	CAD	21840005	2
	rs1800977	upstream variant 2KB, utr variant 5 prime	C/T	A=0.3612/1808	YES	Alzheimer's disease, atherothrombotic cerebral infarction	19446537	4
	rs1800978	intron variant,utr variant 5 prime	C/G	G=0.1418/709	YES	HDLC level	21316679	2
	rs1883025	intron variant	A/G	T=0.3075/1540	YES	AMD	24970616	13
	rs2020927	intron variant	C/T	G=0.3321/1663	YES	Alzheimer's disease	17510949	2
	rs2065412	intron variant	C/T	C=0.1859/931	YES	Dementia	19606474	1
	rs2066714	missense	A/G	C=0.3568/1786	YES	-	-	9
	rs2066718	missense	A/C/G	T=0.0741/371	YES	-	-	4
	rs2230805	synonymous codon	A/G	T=0.3724/1864	YES	Dementia	19606474	3
	rs2230808	missense	A/G	T=0.4617/2311	YES	schizophrenia,AD	21839797 17510949	10
	rs2246298	upstream variant 2KB	C/T	A=0.2202/1103	YES	HDLC level	19041386	1
	rs2246841	synonymous codon	C/T	T=0.1526/764	YES	ischaemic stroke	17553166	1

ABCA1	rs2249891	intron variant	A/G	G=0.3602/1804	YES	CAD	20855565	1	1
	rs2275542	intron variant	C/T	T=0.3053/1529	YES	HDLC level	19041386	1	0
	rs2275543	intron variant	C/T	C=0.0747/373	YES	TC level	20832063	1	1
	rs2275544	intron variant	C/T	C=0.1258/630	YES	Apolipoprotein E levels	17430597	1	0
	rs2297398	intron variant	G/T	G=0.1066/533	YES	-	-	1	0
	rs2297404	intron variant	C/G	G=0.0940/470	YES	Alzheimer's disease	17510949	1	1
	rs2422493	upstream variant 2KB	C/T	A=0.4880/2443	YES	CAD,Alzheimer's disease	21840005 20571217	5	2
	rs2472384	intron variant	A/G	C=0.4441/2223	YES	-	-	1	0
	rs2472386	intron variant	C/T	G=0.4285/2145	YES	Dementia	19606474	1	1
	rs2482424	intron variant	C/T	T=0.1298/649	YES	Alzheimer's disease	20061627	1	0
	rs2515602	intron variant	C/T	A=0.4249/2127	YES	HDLC level	19041386	1	0
	rs2515629	intron variant	C/T	G=0.1408/705	YES	-	-	1	0
	rs2575875	intron variant	A/G	A=0.4938/2473	YES	lipid metabolism	20185793	1	1
	rs2575879	intron variant, upstream variant 2KB	C/G	G=0.3107/1556	YES	HDLC level	21216679	1	1
	rs2740483	upstream variant 2KB	C/G	G=0.2129/1065	YES	-	-	3	0
	rs2740491	intron variant	C/T	G=0.4972/2489	YES	-	-	1	0
	rs2777799	intron variant	A/G	G=0.1138/570	YES	Apolipoprotein E levels	17430597	1	0
	rs3758294	intron variant	C/T	C=0.3003/1503	YES	-	-	1	0
	rs3847303	intron variant	A/G	T=0.1346/674	YES	-	-	2	0
	rs3858075	intron variant	C/T	T=0.2242/1122	YES	schizophrenia	19721717	1	1
	rs3890182	intron variant	A/G	A=0.0982/491	YES	-	-	12	0
	rs3904999	intron variant	A/G	T=0.2288/1146	YES	Apolipoprotein E levels	17430597	1	0
	rs3905000	intron variant	A/G	A=0.1240/621	YES	-	-	8	0
	rs4149263	intron variant	C/T	G=0.1627/815	YES	atherogenic dyslipidaemia	19538231	2	1
	rs4149264	intron variant	C/G	C=0.1613/808	YES	statin responsiveness	21741043	1	1
	rs4149268	intron variant	A/G	T=0.4916/2461	YES	-	-	8	0
	rs4149270	intron variant	A/C/G/T	A=0.3331/1667	YES	-	-	1	0
	rs4149271	intron variant	C/T	A=0.2388/1196	YES	Dementia	19606474	1	1
	rs4149272	intron variant	A/G	C=0.4896/2452	YES	postprandial lipid metabolism	20185793	1	1
	rs4149274	intron variant	C/T	A=0.3341/1672	YES	-	-	1	0
	rs4149310	intron variant	A/T	T=0.4643/2324	YES	-	-	2	0
	rs4149324	intron variant	A/G	C=0.0851/426	YES	-	-	1	0

ABCA1	rs4149327	intron variant	C/G	G=0.2682/1343	YES	-	-	1
	rs4743763	intron variant	A/T	A=0.4525/2265	YES	HDLC level	21347282	1
	rs6479283	intron variant	C/T	T=0.2955/1479	YES	Apolipoprotein E levels	17430597	1
	rs9282541	missense	C/T	A=0.0060/30	YES	HDLC level	20418488 20797885	4
	rs10512338	intron variant	C/T	C=0.1406/703	YES	-	-	2
	rs10820738	intron variant	C/T	C=0.1054/527	YES	-	-	5
	rs12003906	intron variant	A/C/G/T	T=0.0665/332	YES	LDLC level	20031551	1
	rs12350560	intron variant	A/G	A=0.0954/478	YES	Dementia	19606474	1
	rs13284054	intron variant, nc transcript variant, upstream variant 2KB	C/T	C=0.1132/567	YES	-	-	1
	rs33918808	intron variant, missense	C/G	G=0.0567/284	YES	-	-	1
	rs4149311	intron variant	A/G	T=0.3397/1701	YES	-	-	1
	rs2066718	missense	A/C/G	T=0.0741/371	YES	Alzheimer's disease	17324514	4
	rs12686004	intron variant	A/G	A=0.1266/633	YES	-	-	1
	rs2575876	intron variant	A/G	A=0.2632/1318	YES	-	-	1
	rs2740486	intron variant	A/C	T=0.4004/2004	YES	-	-	1
	rs10512337	intron variant,nc transcript variant	A/G	A=0.1000/501	YES	-	-	1
	rs7035693	intron variant	A/G	G=0.1406/703	YES	-	-	1
CP	rs16861582	downstream variant 500B,intron variant	A/G	G=0.3423/1713	YES	-	-	1
	rs6799507	intron variant	A/G	G=0.1292/646	YES	-	-	1
	rs13072552	intron variant	G/T	T=0.1577/790	YES	-	-	1
	rs17838831	upstream variant 2KB	A/G	C=0.2039/1021	YES	-	-	1
C1QA	rs2935542	intron variant	A/G	G=0.1222/611	YES	Lupus	19440201	1
	rs292001	intron variant	A/G	G=0.3504/1754	YES	RA,T2DM	23607884	4
	rs172378	synonymous codon	A/G	A=0.3836/1920	YES	breast cancer,lupus,Familial amyloidotic polyneuropathy	20332777 19493541 19440201	4
	rs12033074	downstream variant 500B	C/G	G=0.4317/2161	YES	Lupus	19440201	1

C1QR1	rs7492	utr variant 3 prime	A/G	A=0.1234/618	YES	renal transplant	22892990	1
	rs2749812	utr variant 3 prime	A/G	A=0.1400/700	YES	CAD	21332844	1
	rs3746731	missense	A/G	G=0.4589/2298	YES	CAD	18599554	2
C1RL	-	-	-	-	-	-	-	-
DDX1	rs2302929	synonymous codon	C/T	C=0.1749/875	YES	dyskeratosis congenita	18252230	1
	rs3755132	upstream variant 2KB	G/T	G=0.3149/1576	YES	Wilms tumor	22544364	1
	rs10221770	utr variant 5 prime	A/G	A=0.2298/1150	YES	dyskeratosis congenita	18252230	1
	rs10929378	synonymous codon	C/T	C=0.2552/1278	YES	dyskeratosis congenita	18252230	1
DMC1	rs5757133	intron variant	A/C/G/T	T=0.2300/1152	YES	cervical cancer	20084279	1
DUT	rs3784621	intron variant	C/T	C=0.4589/2298	YES	cervical cancer	20084279	1
	rs11637235	intron variant	C/T	T=0.3379/1691	YES	-	-	3
ENPP1	rs1510	utr variant 3 prime	C/G	G=0.0521/260	YES	-	-	1
	rs703184	intron variant	C/G	C=0.1064/533	YES	-	-	1
	rs858338	intron variant	G/T	T=0.1296/649	YES	-	-	1
	rs858339	intron variant	A/T	A=0.2809/1406	YES	-	-	1
	rs858341	intron variant	A/G	G=0.4726/2367	YES	-	-	3
	rs858342	intron variant	A/G	G=0.2169/1085	YES	hand osteoarthritis	1257435	2
	rs858345	intron variant	A/G	T=0.4888/2448	YES	-	-	1
	rs943003	intron variant	A/G	C=0.2925/1465	YES	obesity	18551113	1
	rs943004	intron variant	A/G	T=0.0911/455	YES	-	-	1
	rs1044548	utr variant 3 prime	A/G	A=0.1162/581	YES	-	-	1
	rs1409181	intron variant	C/G	G=0.4806/2407	YES	-	-	1
	rs1409182	intron variant	A/C/G/T	T=0.0889/445	YES	-	-	1
	rs1409184	intron variant	C/T	A=0.3462/1733	YES	T2DM	18678618	1
	rs1799774	intron variant	/T	-=0.4762/2384	YES	obesity	18719658	6
	rs1800949	upstream variant 2K	C/T	T=0.1384/693	YES	-	-	2
	rs1830971	intron variant	C/T	G=0.3702/1854	YES	T2DM	18678618	1
	rs1974201	intron variant	C/G	C=0.4109/2057	YES	kidney failure,T2DM	22199358 18184924	3
	rs2021966	intron variant	C/T	G=0.4661/2333	YES	DM	25222839	6
	rs4141767	intron variant	A/G	G=0.1759/881	YES	-	-	1
	rs6569759	intron variant	A/G	A=0.3880/1942	YES	-	-	1
	rs6916495	intron variant	C/T	T=0.1244/623	YES	-	-	1
	rs6917903	intron variant	C/G	C=0.3704/1854	YES	-	-	1

ENPP1	rs7754586	downstream variant 500B,utr variant 3 prime	A/C	C=0.4127/2067	YES	T1DM	18184924	1
	rs7754859	downstream variant 500B,utr variant 3 prime	C/T	C=0.4463/2235	YES	T2DM	18678618	1
	rs7756163	intron variant	C/T	C=0.2881/1443	YES	-	-	1
	rs7767111	intron variant	A/G	A=0.1168/584	YES	-	-	1
	rs7768480	intron variant	A/G	G=0.2326/1165	YES	-	-	1
	rs7773477	intron variant	G/T	T=0.1040/521	YES	Pseudoxanthoma elasticum	25025693	1
	rs7775386	intron variant	C/T	T=0.2542/1272	YES	-	-	1
	rs9372999	intron variant	A/C	A=0.1817/909	YES	-	-	1
	rs9375830	intron variant	A/G	A=0.3606/1805	YES	-	-	1
	rs9402345	intron variant	A/G	A=0.1112/556	YES	-	-	1
	rs9402346	intron variant	C/G	G=0.3460/1732	YES	T2DM	18678618	1
	rs9402348	intron variant	G/T	G=0.2083/1043	YES	-	-	1
	rs9402349	intron variant	A/C	C=0.1348/674	YES	-	-	2
	rs9493105	intron variant	C/T	T=0.1112/556	YES	-	-	1
	rs9493116	intron variant	A/G	G=0.1396/699	YES	-	-	1
	rs9493120	utr variant 3 prime	A/G	A=0.1172/586	YES	-	-	1
	rs9493121	utr variant 3 prime	A/G	G=0.1605/803	YES	-	-	1
	rs11964389	utr variant 3 prime	A/G	G=0.1605/803	YES	obesity	18719658	1
	rs12201710	utr variant 3 prime	C/G	C=0.0789/394	YES	-	-	1
	rs7754561	downstream variant 500B,utr variant 3 prime	A/G	A=0.4453/2230	YES	hypertriglyceridemia,obesity	19656007 18719658	9
EVPL	-	-	-	-	-	-	-	-
GC	rs4752	synonymous codon	C/T	G=0.1056/529	YES	ankylosing spondylitis	21844150	4
	rs221999	intron variant	C/T	G=0.2260/1131	YES	-	-	1
	rs222003	intron variant	C/G	C=0.1877/940	YES	-	-	2
	rs222011	intron variant	G/T	A=0.0599/300	YES	vitamin D metabolism	16600026	1
	rs222014	intron variant	A/G	T=0.0579/289	YES	vitamin D metabolism	23456391	3
	rs222016	intron variant	A/G	G=0.3363/1684	YES	ankylosing spondylitis, vitamin D metabolism	21844150 23456391	3
	rs222020	intron variant	C/T	C=0.3377/1691	YES	vitamin D metabolism, ankylosing spondylitis	22801813 21844150	5

GC	rs222029	intron variant	A/G	G=0.2877/1441	YES	vitamin D metabolism	23505139 19543766	3
	rs222035	intron variant	A/C	G=0.3742/1873	YES	vitamin D metabolism	22144504	2
	rs222040	intron variant	C/T	A=0.4902/2455	YES	vitamin D metabolism	23505139	3
	rs705117	intron variant	A/G	C=0.4217/2111	YES	vitamin D metabolism	23505139	5
	rs705120	intron variant	G/T	A=0.4191/2098	YES	vitamin D metabolism	23505139	2
	rs842999	intron variant	A/C/G	C=0.3578/1791	YES	vitamin D metabolism	23505139	3
	rs843006	intron variant	G/T	T=0.3592/1798	YES	vitamin D metabolism	19255064	1
	rs1155563	intron variant	C/T	C=0.2192/1097	YES	vitamin D metabolism,COPD	23505139 20562129	8
	rs1352843	intron variant	C/T	C=0.0577/289	YES	vitamin D metabolism	18593774	2
	rs1352844	intron variant	C/T	T=0.0851/426	YES	vitamin D metabolism	19255064	1
	rs1491710	intron variant	G/T	C=0.2218/1110	YES	vitamin D metabolism	23505139	1
	rs2070741	intron variant	A/C	G=0.1080/540	YES	COPD	21228423	1
	rs2282679	intron variant	A/C	G=0.2021/1012	YES	vitamin D metabolism,subclinical atherosclerosis,asthma	24868205 24663808 22673963	19
	rs2298849	intron variant	C/T	G=0.2947/1475	YES	vitamin D metabolism	22583563	4
	rs3733359	intron variant, utr variant 5 prime	C/T	A=0.2065/1034	YES	ankylosing spondylitis	21844150	3
	rs10488854	intron variant	A/G	T=0.1004/502	YES	vitamin D metabolism	23505139	1
	rs16846912	intron variant	A/G	G=0.1062/531	YES	vitamin D metabolism	19255064	1
	rs16847015	intron variant	A/C/G/T	A=0.1418/709	YES	vitamin D metabolism	22583563	1
	rs16847024	intron variant, upstream variant 2KB	C/T	T=0.0701/351	YES	colorectal cancer	24562971	1
	rs16847036	intron variant	A/G	G=0.1649/825	YES	vitamin D metabolism	19543766	1
	rs16847054	intron variant	C/T	T=0.1250/626	YES	T2DM	19956108	1
	rs12640179	intron variant	C/G	G=0.1028/514	YES	T2DM	19956108	1
	rs1491709	intron variant	C/T	A=0.1032/517	YES	vitamin D metabolism	18593774	2
GDF10	rs3007	utr variant 3 prime	A/T	A=0.2416/1209	YES	Alzheimer disease	16385451	1
	rs1902725	intron variant	A/G	T=0.2328/1165	YES	-	-	1
GDF2	rs3781226	intron variant	C/T	T=0.0845/423	YES	-	-	1
	rs9971293	intron variant	C/T	C=0.1146/574	YES	Alzheimer disease	16385451	1
GJA1	-	-	-	-	-	-	-	-
GSTCD	-	-	-	-	-	-	-	-
GTF2H4	rs2894054	NIL	C/T	A=0.0855/428	YES	cervical disease	20084279	1
	rs1052693	utr variant 5 prime	C/T	T=0.4077/2042	YES	-	-	1

GTF2H4	rs1264307	intron variant, upstream variant 2KB	C/T	A=0.2220/1112	YES	Aspirin-exacerbated respiratory disease	22524621	1
IFNG	rs11177074	NIL	C/T	C=0.1416/708	YES	cervical disease	20084279	1
	rs1861493	intron variant	A/G	G=0.2095/1049	YES	idiopathic inflammatory myopathy	17405833	6
	rs1861494	intron variant	C/T	C=0.2258/1131	YES	imatinib therapy	20959405	11
	rs2069705	upstream variant 2KB	C/T	G=0.4778/2392	YES	breast cancer,SLE	20418110 19919944	17
	rs2069718	intron variant	C/T	G=0.3832/1919	YES	TB,SLE	23737189 19919944	8
	rs2069727	downstream variant 500B	A/G	C=0.2750/1376	YES	asthma,ALL,MS	21798578 21067287 18332247	11
	rs2430561	intron variant	A/T	A=0.2802/1403	YES	anti-TNF treatment, Henoch-Schonlein purpura,SLE,leprosy resistance,acute renal allograft rejection, vitiligo,ACS,oral lichen planus,bipolar disorder,TB,Chagas disease,SARS	24967817 24659122 24443230 24389160 24057242 23777204 23159874 22041017 21728784 21332391 20359550 16672072	49
IL2RA	rs706778	intron variant	A/G	C=0.4992/2499	YES	UC,alopecia areata,T1DM	23972291 22897480 19106270	12
	rs706779	intron variant	A/G	C=0.4581/2293	YES	-	-	3
	rs706780	intron variant	A/C	A=0.1224/612	YES	-	-	1
	rs791587	intron variant	A/G	A=0.4804/2405	YES	-	-	1
	rs791589	intron variant	A/G	G=0.3576/1791	YES	-	-	3
	rs791590	intron variant	A/T	T=0.1108/555	YES	-	-	1
	rs942201	intron variant	G/T	T=0.1819/911	YES	-	-	1
	rs1107345	intron variant	A/C	T=0.1783/893	YES	-	-	1

IL2RA	rs1323658	intron variant	A/C/G/T	C=0.1498/749	YES	-	-	1
	rs1570538	utr variant 3 prime	C/T	T=0.3760/1882	YES	MS	19125193	2
	rs2025345	intron variant	A/G	T=0.3760/1882	YES	-	-	2
	rs2031229	intron variant	A/G	A=0.1795/899	YES	-	-	1
	rs2076846	intron variant	A/G	G=0.2316/1159	YES	-	-	0
	rs2104286	intron variant	A/G	C=0.1306/653	YES	bladder cancer,MS,RA, giant cell arteritis	24931268 24770783 23529819 20810507	44
	rs2228149	intron variant, synonymous codon	C/T	A=0.0603/301	YES	-	-	3
	rs2228150	synonymous codon	A/G	T=0.0819/409	YES	-	-	1
	rs2256774	intron variant	A/G	T=0.0819/409	YES	MS	22085902	2
	rs2274037	intron variant	A/G	A=0.0603/301	YES	-	-	3
	rs2386841	intron variant	A/C	T=0.2446/1225	YES	-	-	2
	rs2476491	intron variant	A/T	T=0.1909/956	YES	-	-	1
	rs3118470	intron variant	C/T	C=0.3185/1594	YES	T1DM,MS,GD	22211793 22085902 20615141	15
	rs3134883	intron variant	C/T	A=0.2023/1012	YES	T1DM	19956099	3
	rs4749924	intron variant	A/C	C=0.1611/806	YES	UC	23972291	2
	rs6602391	intron variant	A/G	A=0.0905/453	YES	-	-	1
	rs6602398	intron variant	G/T	T=0.1613/808	YES	-	-	1
	rs7072398	intron variant	A/G	G=0.4768/2387	YES	-	-	1
	rs7072793	upstream variant 2KB	C/T	C=0.4511/2258	YES	breast cancer,T1DM	22213266 19956099	6
	rs7093069	intron variant	C/T	T=0.1679/840	YES	T1DM	22211793	2
	rs7899538	intron variant	A/C	A=0.1156/578	YES	-	-	1
	rs7910961	intron variant	C/T	T=0.2272/1137	YES	-	-	1
	rs9663421	intron variant	C/T	T=0.1723/862	YES	-	-	1
	rs10905656	intron variant	A/C	A=0.4397/2201	YES	-	-	1
	rs10905669	intron variant	C/T	T=0.2129/1065	YES	-	-	3
	rs11256448	intron variant	A/G	G=0.2252/1128	YES	-	-	1
	rs11256456	intron variant	C/T	C=0.1899/950	YES	-	-	1

IL2RA	rs11256497	intron variant	A/G	A=0.2139/1071	YES	-	-	2
	rs11596355	intron variant	C/T	C=0.0567/284	YES	-	-	1
	rs12244380	utr variant 3 prime	A/G	G=0.3548/1776	YES	-	-	2
	rs12722489	intron variant	A/G	T=0.0911/455	YES	MS	243332945	16
	rs12722561	intron variant	A/G	T=0.1266/633	YES	-	-	2
	rs12722574	intron variant	C/T	A=0.2482/1243	YES	-	-	1
	rs12727588	intron variant	A/G	T=0.1661/831	YES	-	-	1
	rs12722605	utr variant 3 prime	A/T	A=0.0747/373	YES	-	-	3
	rs12722608	utr variant 3 prime	G/T	A=0.0599/300	YES	-	-	1
ILDR1	rs11718322	intron variant	A/G	A=0.3936/1970	YES	Behçet's disease	20622878	1
LY96	rs11465996	upstream variant 2KB	C/G	G=0.2210/1106	YES	inflammatory bowel disease,sepsis and MOD	24971461 22266968	4
	rs16938755	intron variant	C/T	C=0.1997/1000	YES	glioma	19423540	1
	rs10808798	intron variant	C/T	C=0.4611/2308	YES	wheeze	21389010	1
	rs2114169	intron variant	A/T	T=0.3113/1559	YES	major trauma	22266968	1
MBD3L1	-	-	-	-	-	-	-	-
MICA	rs1882	utr variant 3 prime	A/G	A=0.2768/1385	YES	non-small cell lung cancer	24997136	1
	rs1051790	intron variant, missense	C/G	G=0.2183/1093	YES	Behçet's uveitis	24136464	1
	rs1051792	intron variant, missense	A/G	A=0.3648/1826	YES	SLE	21702010	2
	rs1051794	missense	A/G	A=0.3646/1826	YES	RA	19409079	1
	rs1063632	missense,synonymous codon,utr variant 5 prime	A/G	A=0.0645/322	YES	T1DM	19143814	1
	rs1063635	missense	A/G	G=0.4167/2086	YES	-	-	3
	rs1131896	missense	A/G	A=0.3454/1729	YES	-	-	2
	rs2523451	intron variant, upstream variant 2KB	C/T	A=0.2424/1213	YES	RA	19116921	1
	rs2523454	nc transcript variant, upstream variant 2KB, utr variant 5 prime	C/T	A=0.2005/1003	YES	-	-	1
	rs2596542	intron variant, upstream variant 2KB	A/G	T=0.4215/2111	YES	HCC,HCV,HBV	24357186 24023482 24010643	8

MICA	rs3763288	intron variant, upstream variant 2KB	A/G	A=0.0761/381	YES	RA	19409079	1
	rs7756993	intron variant	C/T	C=0.3433/1719	YES	T1DM	19143813	1
	rs9266825	utr variant 3 prime	A/C	A=0.3429/1716	YES	NSCLC	24997136	1
	rs9380254	missense, utr variant 5 prime	C/G	C=0.0747/373	YES	-	-	1
	rs12660741	intron variant	A/G	G=0.0521/260	YES	leprosy	20617178	1
	rs2857281	intron variant	A/C	C=0.1238/620	YES	Behçet's disease	20622878	1
	rs12175489	intron variant	A/G	A=0.1645/824	YES	viseral fat	22589738	1
	rs2256175	intron variant	A/G	C=0.4105/2055	YES	neuroblastoma	18463370	1
	rs2256183	intron variant	A/G	A=0.2528/1266	YES	height	20881960	1
MOCS2	-	-	-	-	-	-	-	-
NEIL2	rs804268	intron variant	C/G	G=0.3730/1867	YES	colorectal cancer	17029639	1
	rs804269	intron variant	A/G	T=0.3021/1513	YES	colorectal cancer	17029639	1
	rs804270	intron variant, utr variant 5 prime	C/G	G=0.4820/2414	YES	SCCOOP	18594018	1
	rs1534862	utr variant 3 prime	C/T	T=0.2534/1269	YES	colorectal cancer	17029639	1
	rs8191518	upstream variant 2KB, utr variant 5 prime	C/G	G=0.1050/526	YES	-	-	1
	rs8191613	intron variant, missense	A/G	A=0.0739/370	YES	colorectal cancer	17029639	1
	rs8191642	synonymous codon	A/G	G=0.2147/1075	YES	breast cancer	18551366	2
	rs8191649	intron variant	C/T	T=0.1464/733	YES	breast cancer	18551366	1
	rs8191663	intron variant	C/T	T=0.2604/1303	YES	colorectal cancer	17029639	1
	rs8191664	missense	G/T	T=0.0529/264	YES	breast cancer	18551366	2
	rs6601606	intron variant	A/G	G=0.0783/392	YES	HIV	21507922	1
	rs804292	utr variant 3 prime	C/T	G=0.1729/865	YES	-	-	1
OAS1	rs1131454	missense	A/C/G/T	A=0.4738/2373	YES	-	-	10
	rs10774671	intron variant,splice acceptor variant, utr variant 3 prime	A/G	G=0.3856/1930	YES	HFMD,HCV,HBV,T1DM,WNV	25059424 22710942 19799013 19247438 16014697	16

OAS1	rs1051042	intron variant, missense, utr variant 3 prime	C/G/T	G=0.2123/1062	YES	rubella vaccine	20079393	3
	rs1131476	intron variant, missense, utr variant 3 prime	A/C/G/T	G=0.2123/1062	YES	-	-	5
	rs2057778	intron variant	A/C	G=0.2075/1038	YES	West Nile virus infection	19247438	1
	rs2158390	upstream variant 2KB	C/G	C=0.1166/583	YES	West Nile virus infection	19247438	1
	rs2285934	intron variant	A/C	T=0.3500/1753	YES	West Nile virus infection	19247438	2
	rs4766662	intron variant	A/C	A=0.2386/1194	YES	West Nile virus infection	19247438	3
	rs7135577	downstream variant 500B	A/G	A=0.2113/1058	YES	West Nile virus infection	19247438	1
	rs7956880	intron variant	A/C	A=0.1110/555	YES	West Nile virus infection	19247438	1
	rs10744785	intron variant	C/G	C=0.2813/1409	YES	West Nile virus infection	19247438	2
	rs12307655	intron variant	C/T	C=0.1012/506	YES	cervical cancer	20084279	1
	rs34137742	intron variant	C/T	T=0.1030/515	YES	prostate cancer	21638280	1
	rs2660	missense, utr variant 3 prime	A/G	G=0.2123/1062	YES	HCV,prostate cancer	22710942 21638280	9
	rs10744785	intron variant	C/G	C=0.2813/1409	YES	West Nile virus infection	19247438	2
OAS2	rs15895	stop lost, utr variant 3 prime	C/T	A=0.1356/678	YES	Dengue virus,tick-borne encephalitis virus	23337612 21050126	3
	rs1293762	intron variant	A/C	T=0.2288/1146	YES	Dengue virus,tick-borne encephalitis virus	24819159 21050126	3
	rs2072137	intron variant	A/G	C=0.3235/1619	YES	-	-	1
	rs2072138	synonymous codon	C/G	G=0.2190/1097	YES	-	-	1
	rs2240185	intron variant	C/G	C=0.4690/2349	YES	-	-	1
	rs718802	intron variant	G/T	A=0.1416/708	YES	cervical cancer	20084279	1
	rs1732778	intron variant	A/G	A=0.2704/1353	YES	Dengue virus,tick-borne encephalitis virus	23337612 21050126	3
OAS3	rs2010549	utr variant 3 prime	C/G	C=0.2127/1065	YES	-	-	1
	rs2010604	utr variant 3 prime	C/G	G=0.1817/909	YES	HCV	18571276	1
	rs2072134	utr variant 3 prime	C/T	A=0.0242/120	YES	hypertension	24142389	3
	rs2072135	intron variant	A/G	T=0.1961/981	YES	CLL	23668599	1

OAS3	rs2072136	synonymous codon	C/T	A=0.3155/1579	YES	Dengue virus,tick-borne encephalitis virus	23337612 21050126	4
	rs2285932	synonymous codon	A/G	T=0.1575/789	YES	Dengue virus,tick-borne encephalitis virus	23337612 21050126	3
	rs7970983	intron variant	C/T	T=0.3269/1637	YES	RA	20018015	1
	rs10774679	upstream variant 2KB	C/T	C=0.3860/1932	YES	-	-	1
	rs12302655	upstream variant 3KB	A/G	A=0.1256/628	YES	cervical cancer	20084279	1
PGR	rs471767	nc transcript variant, utr variant 3 prime	A/G	G=0.2294/1149	YES	endometrial cancer	19382201	3
	rs480851	intron variant	A/G	G=0.1488/745	YES	-	-	1
	rs484389	nc transcript variant, utr variant 3 prime	C/T	G=0.2618/1310	YES	pelvic organ prolapse	19267271	1
	rs485283	intron variant	C/T	G=0.1542/771	YES	-	-	1
	rs493957	intron variant	A/G	C=0.1540/771	YES	-	-	1
	rs500760	nc transcript variant, synonymous codon	A/G	C=0.2616/1309	YES	endometriosis	16126772	4
	rs506487	intron variant	C/T	T=0.1522/762	YES	-	-	1
	rs508653	intron variant	G/T	A=0.1643/822	YES	-	-	1
	rs518162	nc transcript variant, upstream variant 2KB, utr variant 5 prime	A/G	A=0.1328/665	YES	-	-	3
	rs547378	intron variant	A/G	A=0.2454/1228	YES	endometrial cancer	19382201	1
	rs550778	intron variant	A/C	G=0.1542/771	YES	-	-	1
	rs560291	intron variant	G/T	T=0.1464/733	YES	-	-	1
	rs561650	intron variant	C/T	C=0.0865/433	YES	female reproductive cancer	22121098	1
	rs566351	intron variant	C/T	T=0.4339/2173	YES	breast cancer	16614108	2
	rs572483	intron variant	A/G	C=0.2610/1306	YES	esophageal adenocarcinoma	20453000	1
	rs572698	downstream variant 500B,intron variant	A/C	A=0.2610/1306	YES	-	-	1
	rs578029	intron variant	A/T	A=0.2262/1132	YES	preterm birth	20308837	1
	rs590688	intron variant	C/G	G=0.3798/1901	YES	breast cancer,recurrent pregnancy loss	23764995 21086036	2
	rs596223	intron variant	A/G	G=0.1524/762	YES	-	-	1

PGR	rs608995	nc transcript variant, utr variant 3 prime	A/T	T=0.2488/1246	YES	endometrial cancer,ovarian cancer	20547493 15632380	4
	rs613120	intron variant	C/T	G=0.3900/1952	YES	bladder cancer	19252927	1
	rs619487	intron variant	G/T	G=0.1484/743	YES	-	-	1
	rs660149	intron variant	C/G	G=0.2071/1037	YES	breast cancer	16614108	1
	rs1042838	intron variant, missense, nc transcript variant	G/T	A=0.0715/357	YES	endometrial cancer,RPL, Migraine-associated vertigo,breast cancer	22633539 21086036 17609999 15632380	19
	rs1042839	nc transcript variant, synonymous codon	C/T	A=0.0703/351	YES	endometriosis,ovarian cancer,breast cancer	16126772 15718480 18628428	6
	rs1145463	intron variant	A/T	T=0.1524/762	YES	-	-	1
	rs1379130	intron variant, missense, nc transcript variant, upstream variant 2KB	A/G	A=0.1438/720	YES	-	-	2
	rs3740753	intron variant, missense, nc transcript variant, upstream variant 2KB	C/G	G=0.0721/361	YES	-	-	1
	rs7116336	intron variant	A/T	A=0.3351/1678	YES	breast cancer	16614108	1
	rs10895054	intron variant	A/T	T=0.0699/350	YES	breast cancer	23764995	1
	rs11224561	nc transcript variant, utr variant 3 prime	C/T	T=0.3013/1509	YES	endometrial cancer	21148628	2
	rs11224579	intron variant	C/T	T=0.3181/1592	YES	endometrial cancer	19382201	1
	rs11224592	intron variant	C/T	C=0.2827/1416	YES	RPL	21086036	1
	rs11224598	intron variant	C/T	C=0.2849/1426	YES	endometrial cancer	19382201	1
	rs11571171	intron variant	C/T	G=0.2246/1125	YES	breast cancer	16614108	1
	rs493220	intron variant	A/G	C=0.1486/743	YES	-	-	1
PIK3CA	rs1607237	intron variant	C/T	C=0.3714/1860	YES	breast cancer	22033276	3
	rs2677760	intron variant	C/T	C=0.4856/2431	YES	breast cancer	25108739	2
	rs2699887	intron variant, upstream variant 2KB	A/G	T=0.1176/589	YES	NSCLC,endometrial cancer	24077347 22146979	5

PIK3CA	rs2699905	intron variant	A/G	T=0.1276/639	YES	ovarian cancer	19240718	1
	rs2865084	intron variant, upstream variant 2KB	A/T	T=0.0763/382	YES	ovarian cancer	19240718	1
	rs4855094	intron variant	A/G	A=0.0863/432	YES	esophageal adenocarcinoma	21212151	1
	rs6443624	intron variant	A/C	A=0.2991/1497	YES	sophageal squamous cell carcinoma,endometrial cancer	24945674 22146979	4
	rs7621329	intron variant	C/T	T=0.2702/1353	YES	esophageal squamous cell carcinoma	24945674	3
	rs7640662	intron variant	C/G	G=0.0635/317	YES	-	-	3
	rs7641889	intron variant	C/T	T=0.1012/506	YES	ovarian cancer	19240718	1
	rs7644648	intron variant	A/G	G=0.2702/1353	YES	-	-	3
	rs7651265	intron variant	A/G	G=0.1292/646	YES	esophageal squamous cell carcinoma,colorectal cancer	24945674 20622004	5
	rs9838411	intron variant	A/G	A=0.2368/1185	YES	endometrial cancer	22146979	1
	rs17849071	intron variant	G/T	G=0.1026/514	YES	breast cancer	24908061	1
POLN	rs1923775	intron variant	C/T	C=0.4439/2223	YES	AD	22159054	1
	rs482519	intron variant	C/T	A=0.1967/984	YES	-	-	1
HAUS3	rs17132382	intron variant	C/T	T=0.2440/1222	YES	cervical cancer	20084279	1
PRDX3	rs2271362	intron variant	C/T	T=0.3159/1582	YES	AD	16385451	1
	rs7082598	-	C/T	T=0.1102/552	YES	cervical cancer	22496757	1
PRLH	-	-	-	-	-	-	-	-
RPS19	rs873282	intron variant	A/G	T=0.4433/2220	YES	Diamond-Blackfan Anemia	19587786	1
SULF1	rs437999	intron variant	A/G	A=0.1963/983	YES	cervical cancer	20084279	1
	rs4284050	intron variant	A/C	A=0.4028/2017	YES	cervical cancer	20084279	1
	rs3802278	missense, utr variant 3 prime	A/G	A=0.3317/1660	YES	ovarian cancer	21214932	1
	rs6990375	missense, utr variant 3 prime	A/G	A=0.3125/1565	YES	ovarian cancer	21214932	1
	rs2623047	upstream variant 2KB	C/T	A=0.4736/2371	YES	breast cancer, ovarian cancer	24911625 21214932	2
	rs3087714	utr variant 3 prime	C/T	A=0.3025/1515	YES	ovarian cancer	21214932	1
	rs10108002	intron variant	C/T	T=0.2065/1034	YES	cervical cancer	20084279	1
	rs13264163	utr variant 5 prime	A/G	G=0.1697/849	YES	ovarian cancer	21214932	1
TELO2? RBFOX1	rs4786722	intron variant	C/T	T=0.2945/1475	YES	cervical cancer	22496757	1

TMC6	rs2613514	intron variant	A/G	G=0.1895/949	YES	-	-	1
TMC8	rs7208422	missense,nc transcript variant	A/T	A=0.4477/2242	YES	Squamous cell skin cancer	24913986	5
	rs12452890	missense,nc transcript variant, synonymous codon	A/G	G=0.4597/2302	YES	Squamous cell skin cancer	24913986	1
	rs412611	intron variant	A/G	T=0.0509/254	YES	Squamous cell skin cancer	24913986	1
	rs16970849	downstream variant 500B,intron variant	A/G	A=0.1430/715	YES	cervical disease	21387291	1
	rs454138	utr variant 3 prime	C/G	C=0.3431/1717	YES	schizophrenia	19223858	1
TYMS	rs2790	intron variant, utr variant 3 prime	A/G	G=0.2939/1472	YES	NSCLC,gastric cancer	24997136 24756984	4
	rs502396	intron variant, upstream variant 2KB	C/T	T=0.3998/2001	YES	renal cancer	18098291	5
	rs699517	nc transcript variant, utr variant 3 prime	C/T	T=0.4980/2493	YES	Neural tube defects, lung cancer	22903727 18221821	7
	rs1001761	intron variant	C/T	G=0.3858/1932	YES	-	-	2
	rs1059394	intron variant,nc transcript variant, utr variant 3 prime	A/G	T=0.4978/2493	YES	gastric cancer	24756984	3
	rs2244500	intron variant	C/T	A=0.3840/1923	YES	endometrial cancer	19190136	1
	rs2612100	downstream variant 500B,intron variant, utr variant 3 prime	A/G	G=0.4714/2361	YES	AML	22964418	1
	rs2847149	intron variant	A/G	G=0.3878/1942	YES	-	-	1
	rs2847153	intron variant	A/G	A=0.2901/1453	YES	Neural tube defects	22903727	2
	rs2853532	intron variant, utr variant 3 prime	C/T	T=0.4948/2477	YES	-	-	2
	rs2853533	intron variant, missense	C/G	C=0.3391/1698	YES	colorectal cancer	23893618	3
	rs2853536	intron variant, utr variant 3 prime	C/T	T=0.4978/2493	YES	endometrial cancer	19190136	1
	rs2853539	intron variant, upstream variant 2KB	A/G	A=0.3810/1907	YES	methotrexate treatment	19902562	1

TYMS	rs2853742	intron variant, upstream variant 2KB	C/T	C=0.2222/1112	YES	lung cancer	18221821	1
	rs3786362	synonymous codon	C/T	G=0.0623/311	YES	endometrial cancer	19190136	1
	rs16948305	intron variant	C/T	T=0.0797/398	YES	colorectal cancer	23893618	3
WIF1	rs10506532	intron variant	A/G	T=0.0401/201	YES	-	-	2
CADM1	rs220840	intron variant	C/T	A=0.0891/446	YES	-	17903299	1
	rs314474	intron variant	A/G	T=0.0739/370	YES	-	17903299	1
	rs3802858	intron variant	A/G	C=0.4253/2129	YES	venous thrombosis	19643986	1
	rs4445669	utr variant 3 prime	C/T	C=0.3542/1773	YES	-	-	2
	rs7122693	intron variant	A/G	G=0.2334/1169	YES	-	-	1
	rs7125361	intron variant	C/G	G=0.4954/2480	YES	Alzheimer disease	rs19361613	1
	rs7927241	intron variant	C/G	G=0.4407/2207	YES	venous thrombosis	19643986	1
	rs10790068	intron variant	C/T	C=0.3137/1570	YES	venous thrombosis	19643986	1
	rs10891802	downstream variant 500B,intron variant	A/C	A=0.4083/2045	YES	venous thrombosis	19643986	1
	rs11215397	downstream variant 500B,intron variant	A/G	A=0.3187/1595	YES	venous thrombosis	19643986	1
	rs11215406	intron variant	C/T	C=0.1941/971	YES	-	-	1
	rs11215437	intron variant	C/T	C=0.4203/2104	YES	-	-	1
	rs11215474	intron variant	A/G	G=0.4321/2164	YES	venous thrombosis	19643986	1
	rs220836	intron variant	A/G	C=0.2173/1087	YES	childhood obesity	23251661	1
	rs7950069	intron variant	A/G	A=0.3874/1939	YES	-	-	1
CD83	rs9230	utr variant 3 prime	C/T	T=0.2175/1088	YES	cervical disease	18056445	2
	rs853360	intron variant	C/T	T=0.2354/1178	YES	cervical disease	22134374	3
	rs9296925	intron variant	C/T	C=0.4325/2166	YES	cervical disease	19446866	2
	rs750749	NIL	C/T	C=0.2470/1237	YES	cervical disease	19446866	2
	rs9370729	NIL	C/T	T=0.3187/1595	YES	cervical disease	19446866	2
CDKN2B	rs1063192	intron variant, utr variant 3 prime	C/T	G=0.2053/1028	YES	ESCC,hyperlipidemia,glaucoma,MI	25239644 24103848 22840486 19272367	17
	rs2069416	intron variant, upstream variant 2KB	A/C/T	A=0.3580/1792	YES	ALL	20617153	1

CDKN2B	rs2069418	intron variant, upstream variant 2KB	C/G	G=0.1991/996	YES	-	-	4
	rs2069422	intron variant	A/C	G=0.0639/320	YES	glaucoma	24875940	1
	rs2069426	intron variant	A/C	T=0.0607/304	YES	-	-	2
	rs3217986	intron variant, utr variant 3 prime	A/C	G=0.1168/584	YES	ovatian cancer	19258477	1
	rs3217989	intron variant, utr variant 3 prime	A/G	C=0.0747/373	YES	CAD	21270820	1
	rs3217992	intron variant, utr variant 3 prime	A/G	T=0.3482/1744	YES	POAG,periodontitis,MI	23111177 20978019 19272367	9
	rs10811661	NIL	C/T	C=0.1763/883	YES	impaired glucose metabolism,preeclampsia	20384434 22676277	108
CDKN2A	rs3731201	intron variant	A/G	C=0.0988/495	YES	impaired glucose metabolism	20384434	1
DCLRE1A	rs3750898	missense	C/G	C=0.3423/1713	YES	Alzheimer's disease	16385451	1
FANCD2	rs9849434	intron variant	A/G	A=0.2412/1207	YES	Alzheimer's disease	24755620	1
	rs1552244	intron variant	A/G	G=0.2564/1284	YES	Alzheimer's disease	24755620	1
	rs2272125	intron variant, synonymous codon	A/C	G=0.2422/1212	YES	breast cancer	16679306	1
FASLG	rs6700734	intron variant	A/G	G=0.2576/1290	YES	-	-	1
	rs5030772	intron variant	A/G	G=0.0555/278	YES	-	-	2
FLJ35220	-	-	-	-	-	-	-	-
ITGA4	rs12988934	intron variant	C/T	T=0.1206/603	YES	-	-	1
	rs3770138	intron variant	A/G	T=0.1394/698	YES	RA	25147926	1
	rs2124440	intron variant	C/T	G=0.4692/2349	YES	-	-	3
	rs3770132	intron variant	C/T	G=0.1074/537	YES	-	-	2
	rs155095	intron variant	C/T	A=0.3107/1556	YES	RA	25147926	1
	rs155100	intron variant	A/T	T=0.4497/2252	YES	-	-	3
	rs155101	intron variant	A/G	G=0.3063/1534	YES	-	-	1
	rs155102	intron variant	C/T	A=0.4505/2255	YES	-	-	1
	rs155103	intron variant	C/T	A=0.3059/1532	YES	-	-	1
	rs1038034	intron variant	C/T	G=0.1246/624	YES	-	-	1
	rs2305586	intron variant	A/G	C=0.1406/703	YES	-	-	1
	rs1143674	synonymous codon	A/G	A=0.3385/1694	YES	RA	25147926	2
	rs3770112	intron variant	C/T	A=0.1637/819	YES	-	-	2
	rs1143676	missense	A/G	G=0.2095/1049	YES	MS	19604093	2
	rs1449263	NIL	A/G	C=0.4643/2324	YES	MS	17689671	3

ITGA7	-	-	-	-	-	-	-	-
ILDR1	rs11718322	intron variant	A/G	A=0.3936/1970	YES	Behçet's disease	20622878	1
TNF	rs361525	upstream variant 2KB	A/G	A=0.0609/305	YES	inflammatory bowel disease,end-stage liver disease,psoriasis,cervical cancer,RA,HCC,COPD,GD,breast cancer,juvenile dermatomyositis,asthma	24776844 24361409 24252077 21670964 21385363 21080879 20299531 19732761 19423537 19035492 17450233	88
	rs1799724	downstream variant 500B, upstream variant 2KB	C/T	T=0.0990/495	YES	achalasia,lung cancer,BMD, prostate cancer risk	24259423 24139238 18551993 18196539	52
	rs1799964	downstream variant 500B, upstream variant 2KB	C/T	C=0.2190/1097	YES	HIV,polycystic ovary syndrome,non-tuberculous mycobacteria, dengue hemorrhagic,lung cancer	24896147 23975191 23777929 23380141 22464751	48
	rs1800610	intron variant	C/T	A=0.1004/502	YES	-	-	11
	rs1800629	upstream variant 2KB		A=0.0903/452	YES	TB,thyroid disease,cSSSIs,sepsis,asthma,periodontitis,psoriasis,gastric cancer,prostate cancer,acute kidney injury,breast cancer,NHL,Sjogren's syndrome,cervical disease,HBV,COPD,lung cancer,RA,schizophrenia and bipolar affective disorder	25239251 25127106 25022448 25000179 24936650 24905365 24252077 24142527 24037955 23796916 23530106 22649007 22294627 21670964 19933216 19773451 18713756 18515978	204

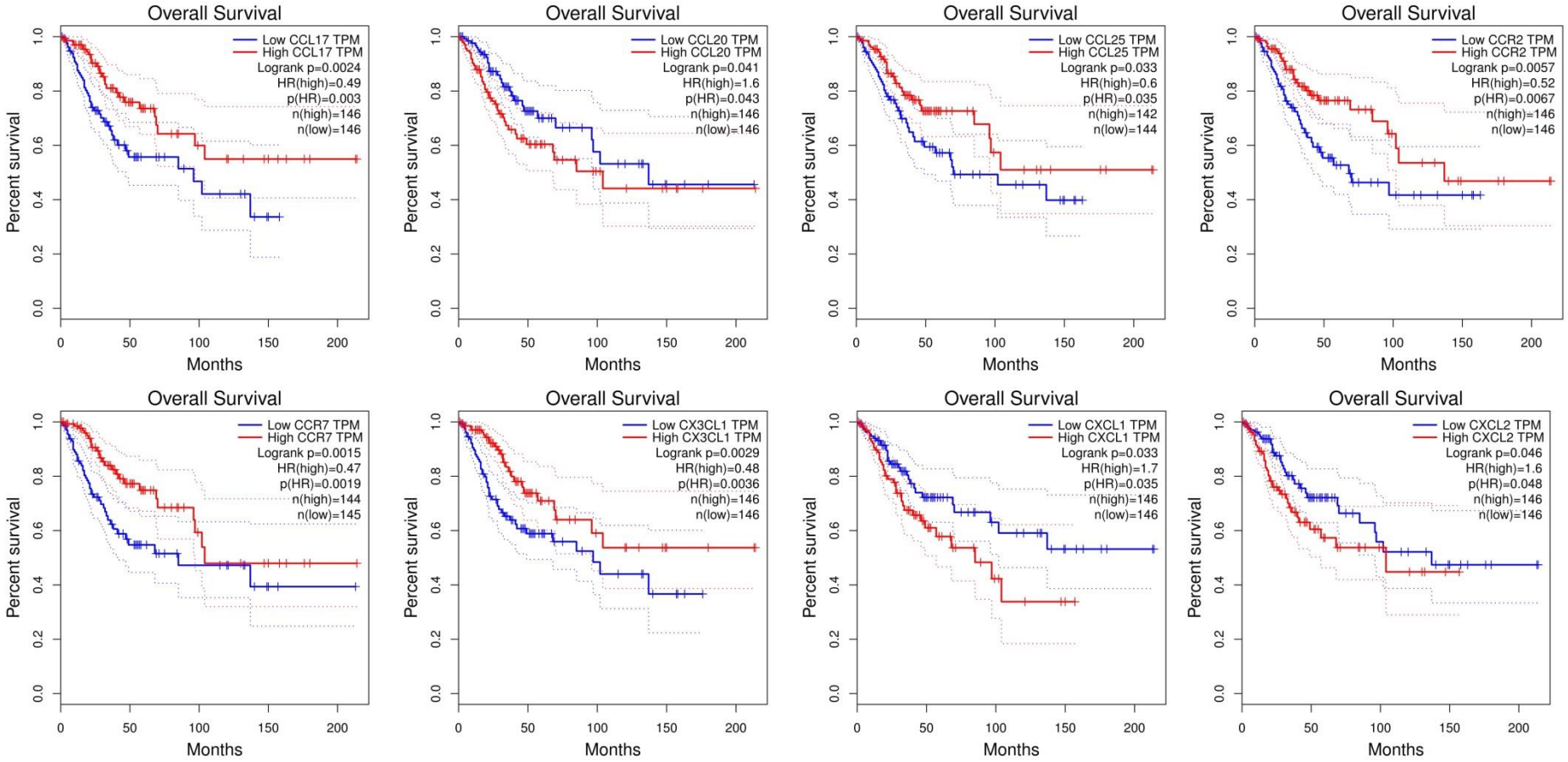
TNF	rs1800630	downstream variant 500B,upstream variant 2KB		A=0.1542/771	YES	recurrent oral ulceration, cardiac surgery,HBV,T2DM, Crohn's disease,gastric cancer,malaria	23406059 21450487 20825556 20177654 19673019 19615068 17493155	38
	rs3093661	intron variant	A/G	A=0.0521/260	YES	Graves' disease	19732761	4
	rs3093662	intron variant	A/G	G=0.0799/399	YES	TB	21463712	14
	rs3093664	intron variant	A/G	G=0.0789/394	YES	-	-	8
	rs3093668	downstream variant 500B,intron variant	C/G	C=0.0511/255	YES	T1DM	19167443	4
WRN	rs1346044	missense	C/T	C=0.1927/965	YES	NHL,Age-related cataract,CKD	25178586 23334603 19282863	8
	rs1800389	synonymous codon	C/T	C=0.3127/1566	YES	-	-	1
	rs1800392	synonymous codon	G/T	T=0.4712/2359	YES	-	-	3
	rs1801195	missense	G/T	T=0.4960/2483	YES	-	-	7
	rs2230009	missense	A/G	A=0.0743/372	YES	breast cancer	17764108	5
	rs3024239	intron variant	C/T	T=0.3750/1878	YES	-	-	1
	rs2553268	intron variant	A/C	G=0.3117/1560	YES	-	-	1
	rs2725338	intron variant	A/G	A=0.1168/584	YES	cataract	22322570	1
	rs2725383	intron variant	C/G	C=0.2400/1202	YES	cataract	22322570	1
	rs3213199	intron variant	A/T	A=0.1004/502	YES	-	-	1
	rs4733220	intron variant	A/G	G=0.4732/2370	YES	cataract	22322570	1
	rs4733225	intron variant	C/T	C=0.1713/858	YES	-	-	1
	rs4987036	intron variant	C/T	T=0.2614/1309	YES	-	-	1
	rs1574311	intron variant	C/T	C=0.1552/777	YES	cataract	22322570	1
	rs13269094	intron variant	G/T	G=0.1585/793	YES	-	-	1
	rs2015230	intron variant	A/G	C=0.3427/1716	YES	-	-	1
	rs2725385	intron variant	A/G	T=0.3081/1542	YES	-	-	1
	rs2725361	intron variant	A/G	G=0.3718/1861	YES	-	-	1
	rs1801195	missense	G/T	T=0.4960/2483	YES	-	-	7
	rs6994361	intron variant	C/T	C=0.4395/2201	YES	-	-	1
	rs2553256	intron variant	A/G	C=0.4191/2098	YES	-	-	1

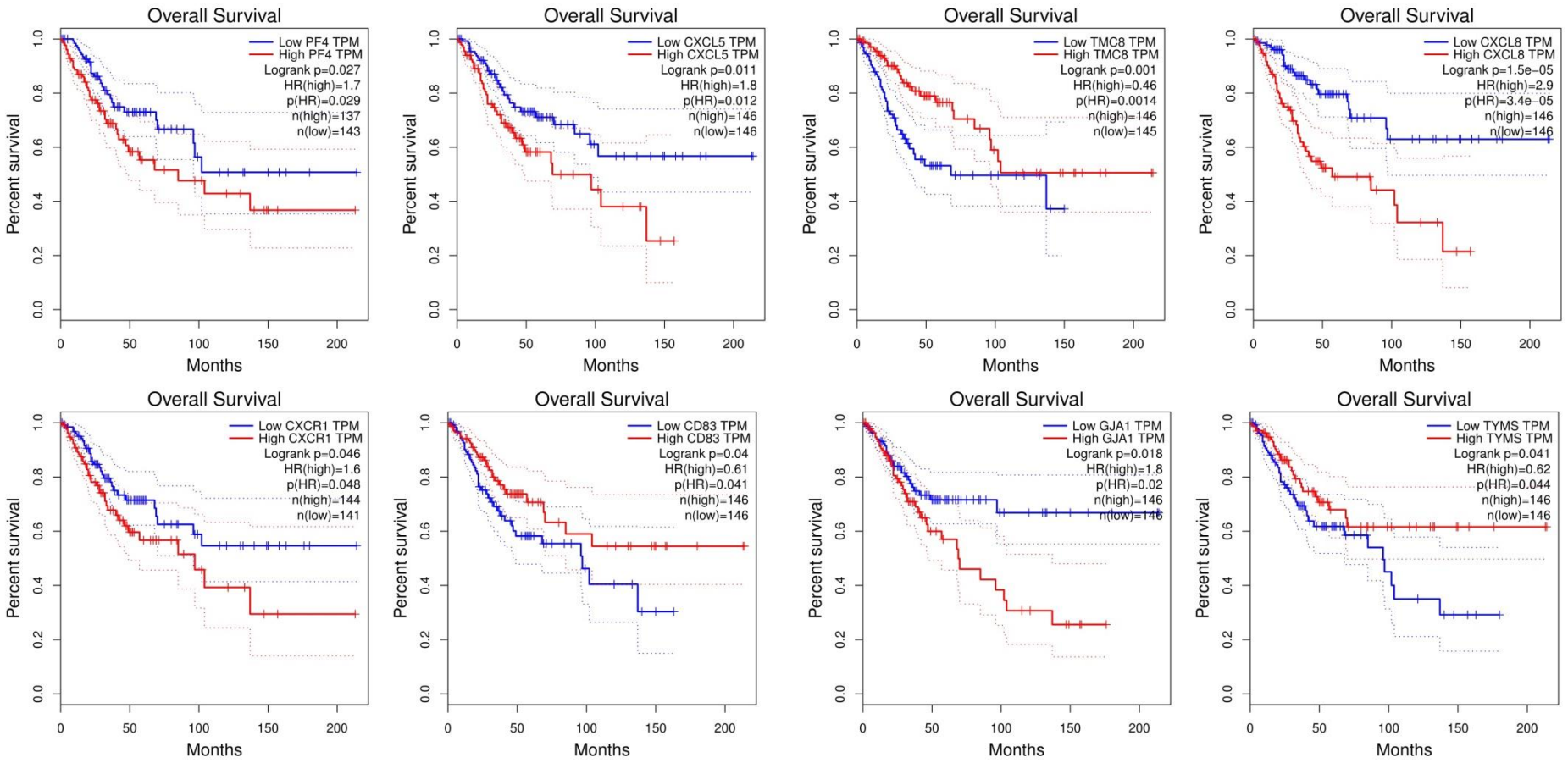
CXCR2	rs1126579	utr variant 3 prime	C/T	T=0.3968/1986	YES	bile duct cancer& biliary stones,bladder cancer, lung cancer	18676870 19252927 25480945	10
	rs2230054	synonymous codon	C/T	C=0.4876/2442	YES	Bevacizumab plus oxaliplatin-based chemotherapy,bile duct cancer& biliary stones,	21791631 18676870	8
	rs1126580	utr variant 3 prime	A/G	A=0.3672/1838	YES	bile duct cancer& biliary stones,diffuse large B-cell lymphoma patients	18676870 18633131	6
	rs4674257	upstream variant 2KB	A/G	A=0.4089/2048	YES	IgA nephropathy	21214373	1
	rs4674258	intron variant, upstream variant 2KB	C/T	T=0.4714/2361	YES	MI	24462138	3
	rs4674259	intron variant, utr variant 5 prime	A/G	G=0.4115/2061	YES	IgA nephropathy	21214373	3
	rs6723449	intron variant	C/T	T=0.3978/1992	YES	MI	24462138	1
	rs11574750	synonymous codon	C/T	T=0.0549/274	YES	UTI	20016852	2
	rs17844697	intron variant	A/G	A=0.3666/1835	YES	HIV	17360650	1
CCR10	rs4683336	intron variant	C/T	C=0.3664/1835	YES	HCV	18822328	1
	rs3919627	downstream variant 500B	C/T	A=0.4385/2195	YES	childhood obesity	23251661	1
CXCR1	rs2234671	missense	C/G	G=0.1442/721	YES	HBV	23396733	10
	rs2671222	upstream variant 2KB	A/G	T=0.1264/633	YES	kidney transplantation	21452410	2
	rs2854386	downstream variant 500B	C/G	G=0.1753/877	YES	cutaneous leishmaniasis	20089160	2
	rs3138060	intron variant	C/G	C=0.0950/475	YES	HIV,UTI	20016852 17360650	2
LIG3	rs3135974	intron variant	A/G	A=0.1012/507	YES	posthematopoietic cell transplant (HCT) outcomes	20226869	2
	rs3135989	intron variant	G/T	G=0.0559/280	YES	-	-	2
	rs2074518	intron variant	A/G	T=0.3105/1555	YES	QT interval	19305408	9
	rs2074517	intron variant	A/G	C=0.0645/323	YES	nasopharyngeal carcinoma	25124928	1
	rs2074516	intron variant	C/G	G=0.2835/1420	YES	-	-	1
	rs2074522	intron variant	C/T	A=0.0789/395	YES	-	-	1

LIG3	rs4796030	intron variant, utr variant 3 prime	A/C	A=0.3840/1923	YES	Recurrent Depressive Disorder, Bladder Cancer treatment	27866211 22701660	5
	rs1052536	downstream variant 500B,utr variant 3 prime	C/T	T=0.3041/1523	YES	Recurrent Depressive Disorder, Fuchs endothelial corneal dystrophy,Colorectal Adenoma, young-onset lung cancer	27866211 25817347 23951112 17108146	8
	rs1003918	downstream variant 500B,utr variant 3 prime	A/G	A=0.3207/1606	YES	Fuchs endothelial corneal dystrophy.	25817347	2
TOPBP1	rs1444601	nc transcript variant, synonymous codon	A/G	G=0.3179/1592	YES	Colorectal cancer	21811255	1
	rs10935070	missense, nc transcript variant	A/C/G	C=0.1879/941	YES	pancreatic cancer	20056645	1
	rs17301766	missense, nc transcript variant	A/G	A=0.0811/406	YES	-	-	1
	rs3192149	missense, nc transcript variant	A/C	T=0.3041/1523	YES	-	-	1
MED24	-	-	-	-	-	-	-	-
CCR1	rs3136671	downstream variant 500B	A/G	T=0.1322/662	YES	-	-	1
	rs3181080	upstream variant 2KB	A/T	T=0.0923/462	YES	-	-	1
XCR1	-	-	-	-	-	-	-	-
CCR2	rs3918357	upstream variant 2KB	A/C	A=0.1659/831	YES	-	-	1
	rs3918358	upstream variant 2KB	A/C	C=0.3285/1645	YES	skeletal muscle strength	20947712 20339010	3
	rs3918359	upstream variant 2KB	A/T	A=0.1947/975	YES	prediabetic condition	18096169	1
	rs3749461	upstream variant 2KB utr variant 5 prime	A/G	G=0.0683/342	YES	-	-	2
	rs3092964	upstream variant 2KB utr variant 5 prime	A/G	G=0.1947/975	YES	-	-	1
	rs3762823	intron variant	A/G	A=0.1947/975	YES	-	-	2

CCR2	rs1799864	missense	G(germline)/A(germline)	A=0.1538/770	YES	recurrent venous thromboembolism, Risk of Carotid Atherosclerosis,migraine,psoriatic arthritis,Kawasaki disease,gastric cancer risk,HIV,prostate cancer,hemophagocytic lymphohistiocytosis, cervical cancer unknown	19263529 19506371 19559392 20153665 17672867 12556692 23632061 25760842	56
	rs1799865	synonymous codon	C/T	C=0.3285/1645	YES	Age-related macular degeneration CK levels	18172114	5
	rs3138042	downstream variant 500B,intron variant, utr variant 3 prime	A/G	G=0.3287/1646	YES	-	-	1
	rs743660	utr variant 3 prime	A/G	A=0.1949/976	YES	IgE levels	19520154	2
DARC	rs13962	missense	A/G	A=0.0689/345	YES	-	22075330	1
	rs2814778	upstream variant 2KB, utr variant 5 prime	A/G	C=0.2664/1334	YES	neutrophil count	-	20

Appendix-4 Kaplan Meier survival plots generated for each gene based on TCGA-CESC database using GEPIA software





To analyze the impact of a specific gene on overall survival of cervical cancer, patient samples in TCGA-CESC were split into two groups according to the median expression of the proposed gene. High expression group (red line) and low expression group (blue line) were compared by a Kaplan-Meier survival plot by both simple log-rank analysis and Cox regression analysis using GEPIA software. Hazard ratio was calculated for each gene based on Cox's proportional hazards Model. 95% confidence intervals were shown as dotted lines. Since the results were only served as supplementary reference standards, raw p values without multiple testing corrections were used. For some genes, the Kaplan-Meier survival curve did not meet the assumptions for log-rank analysis and Cox regression analysis and therefore the p values were not showing the accurate difference

Reference

ABNOVA (2010). Migration assay (https://www.youtube.com/watch?v=6SON7VAA5-k).

Agalliu, I., Gapstur, S., Chen, Z., Wang, T., Anderson, R.L., Teras, L., Kreimer, A.R., Hayes, R.B., Freedman, N.D., and Burk, R.D. (2016). Associations of oral α-, β-, and γ-human papillomavirus types with risk of incident head and neck cancer. JAMA oncology *2*, 599-606.

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control *19*, 716-723.

Aksoy, P., Gottschalk, E.Y., and Meneses, P.I. (2017). HPV entry into cells. Mutation Research/Reviews in Mutation Research *772*, 13-22.

Al-Alwan, L.A., Chang, Y., Mogas, A., Halayko, A.J., Baglole, C.J., Martin, J.G., Rousseau, S., Eidelman, D.H., and Hamid, Q. (2013). Differential roles of CXCL2 and CXCL3 and their receptors in regulating normal and asthmatic airway smooth muscle cell migration. The Journal of Immunology *191*, 2731-2741.

Alfaro, C., Teijeira, A., Oñate, C., Pérez, G., Sanmamed, M.F., Andueza, M.P., Alignani, D., Labiano, S., Azpilikueta, A., and Rodriguez-Paulete, A. (2016). Tumor-produced interleukin-8 attracts human myeloid-derived suppressor cells and elicits extrusion of neutrophil extracellular traps (NETs). Clinical Cancer Research *22*, 3924-3936.

Almajhdi, F.N., Al-Ahdal, M., Abdo, A.A., Sanai, F.M., Al-Anazi, M., Khalaf, N., Viswan, N.A., Al-Ashgar, H., Al-Kahtani, K., and Al-Humaidan, H. (2013). Single nucleotide polymorphisms in CXCR1 gene and its association with hepatitis B infected patients in Saudi Arabia. Annals of hepatology *12*, 220-227.

Anacker, D.C., and Moody, C.A. (2017). Modulation of the DNA damage response during the life cycle of human papillomaviruses. Virus Res *231*, 41-49.

AppliedBiosystems, f. (2010). TaqMan® SNP Genotyping Assays Protocol *Part Number 4332856 Rev. D*.

Araki, S., Omori, Y., Lyn, D., Singh, R.K., Meinbach, D.M., Sandman, Y., Lokeshwar, V.B., and Lokeshwar, B.L. (2007). Interleukin-8 is a molecular determinant of androgen independence and progression in prostate cancer. Cancer research *67*, 6854-6862.

Aul, R., Patel, S., Summerhill, S., Kilty, I., Plumb, J., and Singh, D. (2012). LPS challenge in healthy subjects: an investigation of neutrophil chemotaxis mechanisms involving CXCR1 and CXCR2. International immunopharmacology *13*, 225-231.

Bai, F., Rankinen, T., Charbonneau, C., Belsham, D., Rao, D., Bouchard, C., and Argyropoulos, G. (2004). Functional dimorphism of two hAgRP promoter SNPs in linkage disequilibrium. Journal of medical genetics *41*, 350-353.

Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nature genetics *40*, 340-345.

Barter, E.F., and Stone, M.J. (2012). Synergistic interactions between chemokine receptor elements in recognition of interleukin-8 by soluble receptor mimics. Biochemistry *51*, 1322-1331.

Bates, R.C., DeLeo, M.J., and Mercurio, A.M. (2004). The epithelial–mesenchymal transition of colon carcinoma involves expression of IL-8 and CXCR-1-mediated chemotaxis. Experimental cell research *299*, 315-324.

Beaney, K.E., Cooper, J.A., Drenos, F., and Humphries, S.E. (2017). Assessment of the clinical utility of adding common single nucleotide polymorphism genetic scores to classical risk factor algorithms in coronary heart disease risk prediction in UK men. Clinical Chemistry and Laboratory Medicine (CCLM).

Beck, T., Hastings, R.K., Gollapudi, S., Free, R.C., and Brookes, A.J. (2014). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. European Journal of Human Genetics *22*, 949-952.
Bellamy, R. (2003). Susceptibility to infectious diseases: the importance of host genetics, Vol 4 (Cambridge University Press).
Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological), 289-300.
Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165-1188.
Bodelon, C., Madeleine, M.M., Johnson, L.G., Du, Q., Malkki, M., Petersdorf, E.W., and Schwartz, S.M. (2012). Genetic variation in CD83 and risks of cervical and vulvar cancers: a population-based case-control study. Gynecologic oncology *124*, 525-528.
Boisvert, W.A., Han, X., Tachado, S.D., and Koziel, H. (2012). \(Leu128^{3.43}\)(L128) and\(Val247^{6.40}\)(V247) of CXCR1 Are Critical Amino Acid Residues for G Protein Coupling and Receptor Activation.
Bret, C., Moreaux, J., Schved, J.-F., Hose, D., and Klein, B. (2011). SULFs in human neoplasia: implication as progression and prognosis factors. Journal of translational medicine *9*, 72.
Brookes, A.J. (1999). The essence of SNPs. Gene *234*, 177-186.
Bruni, L., Barrionuevo-Rosas, L., Albero, G., Serrano, B., Mena, M., Gómez, D., Muñoz, J., Bosch, F., and de Sanjosé, S. (2017). Human Papillomavirus and Related Diseases in United Kingdom.Summary Report 27 July 2017. ICO Information Centre on HPV and Cancer (HPV Information Centre).
Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R., Patterson, N., and Robinson, E.B. (2015). An atlas of genetic correlations across human diseases and traits. Nature genetics.
Burnham, K.P., and Anderson, D.R. (2003). Model selection and multimodel inference: a practical information-theoretic approach (Springer Science & Business Media).
Bush, W.S., and Moore, J.H. (2012). Genome-wide association studies. PLoS Comput Biol *8*, e1002822.
Caccuri, F., Giagulli, C., Bugatti, A., Benetti, A., Alessandri, G., Ribatti, D., Marsico, S., Apostoli, P., Slevin, M.A., and Rusnati, M. (2012). HIV-1 matrix protein p17 promotes angiogenesis via chemokine receptors CXCR1 and CXCR2. Proceedings of the National Academy of Sciences *109*, 14580-14585.
Caccuri, F., Rueckert, C., Giagulli, C., Schulze, K., Basta, D., Zicari, S., Marsico, S., Cervi, E., Fiorentini, S., and Slevin, M. (2014). HIV-1 matrix protein p17 promotes lymphangiogenesis and activates the endothelin-1/endothelin B receptor axis. Arteriosclerosis, thrombosis, and vascular biology, ATVBAHA. 113.302478.
Cameron, R.L., Kavanagh, K., Pan, J., Love, J., Cuschieri, K., Robertson, C., Ahmed, S., Palmer, T., and Pollock, K.G. (2016). Human papillomavirus prevalence and herd immunity after introduction of vaccination program, Scotland, 2009–2013. Emerging infectious diseases *22*, 56.
Cardona, S.M., Garcia, J.A., and Cardona, A.E. (2013). The fine balance of chemokines during disease: trafficking, inflammation, and homeostasis. Chemokines: Methods and Protocols, 1-16.
Carrero, Y., Mosquera, J., Callejas, D., and Alvarez-Mon, M. (2015). In situ increased chemokine expression in human cervical intraepithelial neoplasia. Pathology-Research and Practice *211*, 281-285.
Casalicchio, G., Freato, N., Maestri, I., Comar, M., Crovella, S., and Segat, L. (2014). Beta defensin-1 gene polymorphisms and susceptibility to atypical squamous cells of

undetermined significance lesions in Italian gynecological patients. Journal of Medical Virology *86*, 1999-2004.
Castellucci, L., Jamieson, S.E., Miller, E.N., Menezes, E., Oliveira, J., Magalhães, A., Guimarães, L.H., Lessa, M., De Jesus, A.R., and Carvalho, E.M. (2010). CXCR1 and SLC11A1 polymorphisms affect susceptibility to cutaneous leishmaniasis in Brazil: a case-control and family-based study. BMC medical genetics *11*, 10.
Castle, P.E., Schiffman, M., Wheeler, C.M., and Solomon, D. (2009). Evidence for frequent regression of cervical intraepithelial neoplasia-grade 2. Obstetrics and gynecology *113*, 18.
Castro, F.A., Ivansson, E.L., Schmitt, M., Juko-Pecirep, I., Kjellberg, L., Hildesheim, A., Gyllensten, U.B., and Pawlita, M. (2012). Contribution of TMC6 and TMC8 (EVER1 and EVER2) variants to cervical cancer susceptibility. International Journal of Cancer *130*, 349-355.
Catarinoa, A.C.A.M.R., Sousaa, D.P.H., and Medeirosa, D.P.R. (2007). The influence of chemokine receptor CCR2 genotypes in the route to cervical carcinogenesis. Gynecol Obstet Invest *66*, 2.
Chakrabarti, S., and Patel, K.D. (2005). Regulation of matrix metalloproteinase-9 release from IL-8-stimulated human neutrophils. Journal of leukocyte biology *78*, 279-288.
Chatterjee, K., Dandara, C., Hoffman, M., and Williamson, A.-L. (2010). CCR2-V64I polymorphism is associated with increased risk of cervical cancer but not with HPV infection or pre-cancerous lesions in African women. BMC cancer *10*, 278.
Cheli, S., Pietrantonio, F., Clementi, E., and Falvella, F.S. (2015). LightSNiP assay is a good strategy for pharmacogenetics test. Frontiers in pharmacology *6*.
Chen, D., Cui, T., Ek, W.E., Liu, H., Wang, H., and Gyllensten, U. (2015). Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability. Carcinogenesis, bgv083.
Chen, D., Enroth, S., Ivansson, E., and Gyllensten, U. (2014a). Pathway analysis of cervical cancer genome-wide association study highlights the MHC region and pathways involved in response to infection. Human molecular genetics *23*, 6047-6060.
Chen, D., Enroth, S., Liu, H., Sun, Y., Wang, H., Yu, M., Deng, L., Xu, S., and Gyllensten, U. (2016). Pooled analysis of genome-wide association studies of cervical intraepithelial neoplasia 3 (CIN3) identifies a new susceptibility locus. Oncotarget *7*, 42216.
Chen, D., and Gyllensten, U. (2014). A cis-eQTL of HLA-DRB1 and a frameshift mutation of MICA contribute to the pattern of association of HLA alleles with cervical cancer. Cancer Medicine *3*, 445-452.
Chen, D., and Gyllensten, U. (2015). Lessons and implications from association studies and post-GWAS analyses of cervical cancer. Trends in Genetics *31*, 41-54.
Chen, D., Hammer, J., Lindquist, D., Idahl, A., and Gyllensten, U. (2014b). A variant upstream of HLA-DRB1 and multiple variants in MICA influence susceptibility to cervical cancer in a Swedish population. Cancer Medicine *3*, 190-198.
Chen, D., Juko-Pecirep, I., Hammer, J., Ivansson, E., Enroth, S., Gustavsson, I., Feuk, L., Magnusson, P.K., McKay, J.D., and Wilander, E. (2013). Genome-wide association study of susceptibility loci for cervical cancer. Journal of the National Cancer Institute *105*, 624-633.
Chen, L., Fan, J., Chen, H., Meng, Z., Chen, Z., Wang, P., and Liu, L. (2014c). The IL-8/CXCR1 axis is associated with cancer stem cell-like properties and correlates with clinical prognosis in human pancreatic cancer cases. Scientific reports *4*.
Chen, Z., Fan, J.Q., Li, J., Li, Q.S., Yan, Z., Jia, X.K., Liu, W.D., Wei, L.J., Zhang, F.Z., and Gao, H. (2009). Promoter hypermethylation correlates with the Hsulf‐1 silencing in human breast and gastric cancer. International journal of cancer *124*, 739-744.

Cheng, Y.-M., Chen, M.-J., Chen, Y., and Hsu, Y.-C. (2014). Regulation of metastatic action through CXCR4 and HPV16 E6/7 in cervical carcinogenesis. Cancer Cell & Microenvironment *1*.
Chuntharapai, A., and Kim, K.J. (1995). Regulation of the expression of IL-8 receptor A/B by IL-8: possible functions of each receptor. The Journal of Immunology *155*, 2587-2594.
Coelho, A., Matos, A., Catarino, R., Pinto, D., Pereira, D., Lopes, C., and Medeiros, R. (2005). Protective role of the polymorphism CCR2-64I in the progression from squamous intraepithelial lesions to invasive cervical carcinoma. Gynecologic oncology *96*, 760-764.
Cohen-Hillel, E., Mintz, R., Meshel, T., Garty, B.-Z., and Ben-Baruch, A. (2009). Cell migration to the chemokine CXCL8: paxillin is activated and regulates adhesion and cell motility. Cellular and molecular life sciences *66*, 884-899.
Cohen-Hillel, E., Yron, I., Meshel, T., Soria, G., Attal, H., and Ben-Baruch, A. (2006). CXCL8-induced FAK phosphorylation via CXCR1 and CXCR2: cytoskeleton-and integrin-related mechanisms converge with FAK regulatory pathways in a receptor-specific manner. Cytokine *33*, 1-16.
Colacino-Silva, F., de Oliveira Kleine, J.P.F., and Salzgeber, M.B. (2017). Polymorphic DNA repair genes XRCC1 and XRCC3 and the risk for cervical cancer in Brazilian patients. Braz J Oncol *13*, 1-8.
Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. New England Journal of Medicine *372*, 793-795.
Consortium, G.P. (2015). A global reference for human genetic variation. Nature *526*, 68-74.
Crews, K.R., Gaedigk, A., Dunnenberger, H.M., Leeder, J.S., Klein, T.E., Caudle, K.E., Haidar, C.E., Shen, D.D., Callaghan, J.T., and Sadhasivam, S. (2014). Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. Clinical Pharmacology & Therapeutics *95*, 376-382.
Cuschieri, K., Kavanagh, K., Moore, C., Bhatia, R., Love, J., and Pollock, K.G. (2016). Impact of partial bivalent HPV vaccination on vaccine-type infection: a population-based analysis. British journal of cancer *114*, 1261.
Czene, K., Lichtenstein, P., and Hemminki, K. (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family‐cancer database. International Journal of Cancer *99*, 260-266.
David, J.M., Dominguez, C., Hamilton, D.H., and Palena, C. (2016). The IL-8/IL-8R Axis: A Double Agent in Tumor Immune Resistance. Vaccines *4*, 22.
de Abreu, A.L., Malaguti, N., Souza, R.P., Uchimura, N.S., Ferreira, É.C., Pereira, M.W., Carvalho, M.D., Pelloso, S.M., Bonini, M.G., and Gimenes, F. (2016). Association of human papillomavirus, Neisseria gonorrhoeae and Chlamydia trachomatis co-infections on the risk of high-grade squamous intraepithelial cervical lesion. American journal of cancer research *6*, 1371.
de Haan, H.G., Bezemer, I.D., Doggen, C.J., Le Cessie, S., Reitsma, P.H., Arellano, A.R., Tong, C.H., Devlin, J.J., Bare, L.A., and Rosendaal, F.R. (2012). Multiple SNP testing improves risk prediction of first venous thrombosis. Blood *120*, 656-663.
de Haas, S., Delmar, P., Bansal, A.T., Moisse, M., Miles, D.W., Leighl, N., Escudier, B., Van Cutsem, E., Carmeliet, P., and Scherer, S.J. (2014). Genetic variability of VEGF pathway genes in six randomized phase III trials assessing the addition of bevacizumab to standard therapy. Angiogenesis *17*, 909-920.
de Martel, C., Plummer, M., Vignat, J., and Franceschi, S. (2017). Worldwide burden of cancer attributable to HPV by site, country and HPV type. International Journal of Cancer.
De Villiers, E.-M., Fauquet, C., Broker, T.R., Bernard, H.-U., and zur Hausen, H. (2004). Classification of papillomaviruses. Virology *324*, 17-27.

DiGiuseppe, S., Bienkowska-Haba, M., Guion, L.G., and Sapp, M. (2017). Cruising the cellular highways: How human papillomavirus travels from the surface to the nucleus. Virus research *231*, 1-9.
Ding, H., Cai, J., Mao, M., Fang, Y., Huang, Z., Jia, J., Li, T., Xu, L., Wang, J., and Zhou, J. (2014). Tumor‐associated macrophages induce lymphangiogenesis in cervical cancer via interaction with tumor cells. APmIs *122*, 1059-1069.
Doorbar, J. (2013). The E4 protein; structure, function and patterns of expression. Virology *445*, 80-98.
Doorbar, J., Egawa, N., Griffin, H., Kranjec, C., and Murakami, I. (2015). Human papillomavirus molecular biology and disease association. Reviews in medical virology *25*, 2-23.
Dorak, M.T. (2016). Genetic Association Studies: Background, Conduct, Analysis, Interpretation (Garland Science).
Dudbridge, F., and Koeleman, B.P. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. The American Journal of Human Genetics *75*, 424-435.
Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association *50*, 1096-1121.
Dunstan, C.-A.N., Salafranca, M.N., Adhikari, S., Xia, Y., Feng, L., and Harrison, J.K. (1996). Identification of two rat genes orthologous to the human interleukin-8 receptors. Journal of Biological Chemistry *271*, 32770-32776.
Eikawa, S., Ohue, Y., Kitaoka, K., Aji, T., Uenaka, A., Oka, M., and Nakayama, E. (2010). Enrichment of Foxp3+ CD4 regulatory T cells in migrated T cells to IL-6–and IL-8–expressing tumors through predominant induction of CXCR1 by IL-6. The Journal of Immunology *185*, 6734-6740.
England, G. (2016). The 100,000 genomes project. The *100*, 0-2.
Evans, D., Brentnall, A., Byers, H., Harkness, E., Stavrinos, P., Howell, A., Newman, W., and Cuzick, J. (2016). Use of multiple Single Nucleotide Polymorphism (SNP) testing to predict breast cancer risk in a familial screening clinic. Journal of Medical Genetics.
Famooto, A., Almujtaba, M., Dareng, E., Akarolo-Anthony, S., Ogbonna, C., Offiong, R., Olaniyan, O., Wheeler, C.M., Doumatey, A., and Rotimi, C.N. (2013). RPS19 and TYMS SNPs and prevalent high risk human papilloma virus infection in Nigerian women. PloS one *8*, e66930.
Fan, X., Patera, A.C., Pong-Kennedy, A., Deno, G., Gonsiorek, W., Manfra, D.J., Vassileva, G., Zeng, M., Jackson, C., and Sullivan, L. (2007). Murine CXCR1 is a functional receptor for GCP-2/CXCL6 and interleukin-8/CXCL8. Journal of Biological Chemistry *282*, 11658-11666.
Feniger-Barish, R., Yron, I., Meshel, T., Matityahu, E., and Ben-Baruch, A. (2003). IL-8-induced migratory responses through CXCR1 and CXCR2: association with phosphorylation and cellular redistribution of focal adhesion kinase. Biochemistry *42*, 2874-2886.
Fischer, M., Uxa, S., Stanko, C., Magin, T.M., and Engeland, K. (2017). Human papilloma virus E7 oncoprotein abrogates the p53-p21-DREAM pathway. *7*, 2603.
Fu, W., Zhang, Y., Zhang, J., and Chen, W.-F. (2005). Cloning and characterization of mouse homolog of the CXC chemokine receptor CXCR1. Cytokine *31*, 9-17.
Fujimoto, J., Sakaguchi, H., Aoki, I., and Tamaya, T. (2000). Clinical implications of expression of interleukin 8 related to angiogenesis in uterine cervical cancers. Cancer research *60*, 2632-2635.

Galarneau, G., Palmer, C.D., Sankaran, V.G., Orkin, S.H., Hirschhorn, J.N., and Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nature genetics *42*, 1049-1051.
Gallup, D.G. (2008). The spread and staging of cervical cancer. Glob Libr Women's Med (ISSN: 1756-2228).
Gauderman, W. (2006). QUANTO 1.2: A computer program for power and sample size calculation for genetic-epidemiology studies. http://hydra usc edu/gxe.
Gerberding, J.L. (2004). Report to congress. Prevention of Genital Human Papillomavirus Infection Atlanta7 US Centers for Disease Control and Prevention.
Gerger, A., El-Khoueiry, A., Zhang, W., Yang, D., Singh, H., Bohanes, P., Ning, Y., Winder, T., LaBonte, M.J., and Wilson, P.M. (2011). Pharmacogenetic angiogenesis profiling for first-line bevacizumab plus oxaliplatin-based chemotherapy in patients with metastatic colorectal cancer. Clinical Cancer Research *17*, 5783-5792.
Giagulli, C., Magiera, A.K., Bugatti, A., Caccuri, F., Marsico, S., Rusnati, M., Vermi, W., Fiorentini, S., and Caruso, A. (2012). HIV-1 matrix protein p17 binds to the IL-8 receptor CXCR1 and shows IL-8–like chemokine activity on monocytes through Rho/ROCK activation. Blood *119*, 2274-2283.
Gilli, U.O., Schneider, M.K., Loetscher, P., and Seebach, J.D. (2005). Human polymorphonuclear neutrophils are recruited by porcine chemokines acting on CXC chemokine receptor 2, and platelet-activating factor. Transplantation *79*, 1324-1331.
Ginestier, C., Liu, S., Diebel, M.E., Korkaya, H., Luo, M., Brown, M., Wicinski, J., Cabaud, O., Charafe-Jauffret, E., and Birnbaum, D. (2010). CXCR1 blockade selectively targets human breast cancer stem cells in vitro and in xenografts. The Journal of clinical investigation *120*, 485.
Glynn, P., Henney, E., and Hall, I. (2002). The selective CXCR2 antagonist SB272844 blocks interleukin-8 and growth-related oncogene-α-mediated inhibition of spontaneous neutrophil apoptosis. Pulmonary pharmacology & therapeutics *15*, 103-110.
González, J.R., Armengol, L., Solé, X., Guinó, E., Mercader, J.M., Estivill, X., and Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. Bioinformatics *23*, 654-655.
Govindaraju, V., Michoud, M.-C., Al-Chalabi, M., Ferraro, P., Powell, W.S., and Martin, J.G. (2006). Interleukin-8: novel roles in human airway smooth muscle cell contraction and migration. American Journal of Physiology-Cell Physiology *291*, C957-C965.
Graham, S.V. (2010). Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies. Future microbiology *5*, 1493-1506.
Griffith, J.W., Sokol, C.L., and Luster, A.D. (2014). Chemokines and chemokine receptors: positioning cells for host defense and immunity. Annual review of immunology *32*, 659-702.
Håkansson, G., Lutay, N., Andersson, M., Hallgren, O., Westergren-Thorsson, G., Svensson, M., and Godaly, G. (2013). Epithelial G protein-coupled receptor kinases regulate the initial inflammatory response during mycobacterial infection. Immunobiology *218*, 984-994.
Ha, H., Debnath, B., and Neamati, N. (2017). Role of the CXCL8-CXCR1/2 Axis in Cancer and Inflammatory Diseases. Theranostics *7*, 1543.
Ha, N.-T., Freytag, S., and Bickeboeller, H. (2014). Coverage and efficiency in current SNP chips. European Journal of Human Genetics *22*, 1124-1130.
Han, C.H., Huang, Y.-J., Lu, K.H., Liu, Z., Mills, G.B., Wei, Q., and Wang, L.-E. (2011). Polymorphisms in the SULF1 gene are associated with early age of onset and survival of ovarian cancer. Journal of Experimental & Clinical Cancer Research *30*, 5.
Han, X.-g., Du, L., Qiao, H., Tu, B., Wang, Y.-g., Qin, A., Dai, K.-r., Fan, Q.-m., and Tang, T.-t. (2015). CXCR1 knockdown improves the sensitivity of osteosarcoma to cisplatin. Cancer letters *369*, 405-415.

Han, X., Tachado, S.D., Koziel, H., and Boisvert, W.A. (2012). Leu1283. 43 (L128) and Val2476. 40 (V247) of CXCR1 are critical amino acid residues for G protein coupling and receptor activation. PloS one *7*, e42765.
Hartl, D., Latzin, P., Hordijk, P., Marcos, V., Rudolph, C., Woischnik, M., Krauss-Etschmann, S., Koller, B., Reinhardt, D., and Roscher, A.A. (2007). Cleavage of CXCR1 on neutrophils disables bacterial killing in cystic fibrosis lung disease. Nature medicine *13*, 1423.
Hawn, T.R., Scholes, D., Wang, H., Li, S.S., Stapleton, A.E., Janer, M., Aderem, A., Stamm, W.E., Zhao, L.P., and Hooton, T.M. (2009). Genetic variation of the human urinary tract innate immune response and asymptomatic bacteriuria in women. PLoS One *4*, e8300.
Hershfield, M., Callaghan, J., Tassaneeyakul, W., Mushiroda, T., Thorn, C., Klein, T., and Lee, M. (2013). Clinical Pharmacogenetics Implementation Consortium Guidelines for Human Leukocyte Antigen‐B Genotype and Allopurinol Dosing. Clinical Pharmacology & Therapeutics *93*, 153-158.
Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences *106*, 9362-9367.
Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika *75*, 800-802.
Hollingworth, R., and Grand, R.J. (2015). Modulation of DNA damage and repair pathways by human tumour viruses. Viruses *7*, 2542-2591.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 65-70.
Holmes, W.E., Lee, J., Kuang, W.-J., Rice, G.C., and Wood, W.I. (1991). Structure and functional expression of a human interleukin-8 receptor. Science, 1278-1280.
Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika *75*, 383-386.
Horita, N., and Kaneko, T. (2015). Genetic model selection for a case–control study and a meta-analysis. Meta gene *5*, 1-8.
Horvath, C.A., Boulet, G.A., Renoux, V.M., Delvenne, P.O., and Bogers, J.-P.J. (2010). Mechanisms of cell entry by human papillomaviruses: an overview. Virology journal *7*, 11.
Hsing, A.W., Sakoda, L.C., Rashid, A., Andreotti, G., Chen, J., Wang, B.-S., Shen, M.-C., Chen, B.E., Rosenberg, P.S., and Zhang, M. (2008). Variants in inflammation genes and the risk of biliary tract cancers and stones: a population-based study in China. Cancer research *68*, 6442-6452.
Huang, J., Yao, J.L., Zhang, L., Bourne, P.A., Quinn, A.M., di Sant'Agnese, P.A., and Reeder, J.E. (2005). Differential expression of interleukin-8 and its receptors in the neuroendocrine and non-neuroendocrine compartments of prostate cancer. The American journal of pathology *166*, 1807-1815.
Huang, Y., Zhang, J., Cui, Z.M., Zhao, J., and Zheng, Y. (2013). Expression of the CXCL12/CXCR4 and CXCL16/CXCR6 axes in cervical intraepithelial neoplasia and cervical cancer. Chinese journal of cancer *32*, 289-296.
Iftner, T., Elbel, M., Schopp, B., Hiller, T., Loizou, J.I., Caldecott, K.W., and Stubenrauch, F. (2002). Interference of papillomavirus E6 protein with single-strand break repair by interaction with XRCC1. The EMBO journal *21*, 4741-4748.
Inamura, H., Kurosawa, M., Okano, A., Kayaba, H., and Majima, M. (2002). Expression of the interleukin-8 receptors CXCR1 and CXCR2 on cord-blood-derived cultured human mast cells. International archives of allergy and immunology *128*, 142.

Infanger, D.W., Cho, Y., Lopez, B.S., Mohanan, S., Liu, S.C., Gursel, D., Boockvar, J.A., and Fischbach, C. (2013). Glioblastoma stem cells are regulated by interleukin-8 signaling in a tumoral perivascular niche. Cancer research *73*, 7079-7089.
Ivansson, E.L., Gustavsson, I.M., Magnusson, J.J., Steiner, L.L., Magnusson, P.K., Erlich, H.A., and Gyllensten, U.B. (2007). Variants of chemokine receptor 2 and interleukin 4 receptor, but not interleukin 10 or Fas ligand, increase risk of cervical cancer. International Journal of Cancer *121*, 2451-2457.
Jia, M., Han, J., Hang, D., Jiang, J., Wang, M., Wei, B., Dai, J., Zhang, K., Guo, L., and Qi, J. (2016). HLA-DP is the cervical cancer susceptibility loci among women infected by high-risk human papillomavirus: potential implication for triage of human papillomavirus-positive women. Tumor Biology *37*, 8019-8025.
Johanneson, B., Chen, D., Enroth, S., Cui, T., and Gyllensten, U. (2014). Systematic validation of hypothesis-driven candidate genes for cervical cancer in a genome-wide association study. Carcinogenesis *35*, 2084-2088.
Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'donnell, C.J., and De Bakker, P.I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics *24*, 2938-2939.
Jones, S.A., Dewald, B., Clark-Lewis, I., and Baggiolini, M. (1997). Chemokine Antagonists That Discriminate between Interleukin-8 Receptors SELECTIVE BLOCKERS OF CXCR2. Journal of Biological Chemistry *272*, 16166-16169.
Jones, S.A., Wolf, M., Qin, S., Mackay, C.R., and Baggiolini, M. (1996). Different functions for the interleukin 8 receptors (IL-8R) of human neutrophil leukocytes: NADPH oxidase and phospholipase D are activated through IL-8R1 but not IL-8R2. Proceedings of the National Academy of Sciences *93*, 6682-6686.
Joseph, P.R.B., Sarmiento, J.M., Mishra, A.K., Das, S.T., Garofalo, R.P., Navarro, J., and Rajarathnam, K. (2010). Probing the role of CXC motif in chemokine CXCL8 for high affinity binding and activation of CXCR1 and CXCR2 receptors. Journal of Biological Chemistry *285*, 29262-29269.
Kamalakar, A., Bendre, M.S., Washam, C.L., Fowler, T.W., Carver, A., Dilley, J.D., Bracey, J.W., Akel, N.S., Margulies, A.G., and Skinner, R.A. (2014). Circulating interleukin-8 levels explain breast cancer osteolysis in mice and humans. Bone *61*, 176-185.
Kapeu, A.S., Luostarinen, T., Jellum, E., Dillner, J., Hakama, M., Koskela, P., Lenner, P., Löve, A., Mahlamaki, E., and Thoresen, S. (2008). Is smoking an independent risk factor for invasive cervical cancer? A nested case-control study within Nordic biobanks. American journal of epidemiology *169*, 480-488.
Karki, R., Pandya, D., Elston, R.C., and Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. BMC medical genomics *8*, 37.
Kato, H., Tsuchiya, N., and Tokunaga, K. (2000). Single nucleotide polymorphisms in the coding regions of human CXC-chemokine receptors CXCR1, CXCR2 and CXCR3. Genes and immunity *1*, 330.
Khan, M.N., Wang, B., Wei, J., Zhang, Y., Li, Q., Luan, X., Cheng, J.-W., Gordon, J.R., Li, F., and Liu, H. (2015). CXCR1/2 antagonism with CXCL8/interleukin-8 analogue CXCL8 (3–72) K11R/G31P restricts lung cancer growth by inhibiting tumor cell proliferation and suppressing angiogenesis. Oncotarget *6*, 21315.
Kim, J., Chan, C., Elwell, C., Singer, M., Dierks, T., Lemjabbar‐Alaoui, H., Rosen, S., and Engel, J. (2013). Endosulfatases SULF1 and SULF2 limit Chlamydia muridarum infection. Cellular microbiology *15*, 1560-1571.

Kines, R.C., Cerio, R.J., Roberts, J.N., Thompson, C.D., de Los Pinos, E., Lowy, D.R., and Schiller, J.T. (2016). Human papillomavirus capsids preferentially bind and infect tumor cells. International Journal of Cancer *138*, 901-911.
King, C.R., and Nicolae, D.L. (2014). GWAS to sequencing: divergence in study design and analysis. Genes *5*, 460-476.
Kredel, S., Wolff, M., Gierschik, P., and Heilker, R. (2014). Phenotypic Analysis of Chemokine-Driven Actin Reorganization in Primary Human Neutrophils. Assay and drug development technologies *12*, 120-128.
Kruglyak, L., and Nickerson, D.A. (2001). Variation is the spice of life. Nature genetics *27*, 234-235.
Kryczek, I., Wei, S., Keller, E., Liu, R., and Zou, W. (2007). Stroma-derived factor (SDF-1/CXCL12) and human tumor pathogenesis. American Journal of Physiology-Cell Physiology *292*, C987-C995.
Kwon, J.M., and Goate, A.M. (2000). The candidate gene approach. Alcohol research and health *24*, 164-168.
Lütschg, V., Boucke, K., Hemmi, S., and Greber, U.F. (2011). Chemotactic antiviral cytokines promote infectious apical entry of human adenovirus into polarized epithelial cells. Nature communications *2*, 391.
Lai, J.-P., Sandhu, D.S., Shire, A.M., and Roberts, L.R. (2008). The tumor suppressor function of human sulfatase 1 (SULF1) in carcinogenesis. Journal of gastrointestinal cancer *39*, 149-158.
Lane, B.R., Lore, K., Bock, P.J., Andersson, J., Coffey, M.J., Strieter, R.M., and Markovitz, D.M. (2001). Interleukin-8 stimulates human immunodeficiency virus type 1 replication and is a potential new target for antiretroviral therapy. Journal of virology *75*, 8195-8202.
Lebrecht, A., Hefler, L., Tempfer, C., and Koelbl, H. (2001). Serum cytokine concentrations in patients with cervical cancer: interleukin-4, interferon-gamma, and monocyte chemoattractant protein-1. Gynecologic oncology *83*, 170-171.
Leckband, S., Kelsoe, J., Dunnenberger, H., George, A., Tran, E., Berger, R., Müller, D., Whirl‐Carrillo, M., Caudle, K., and Pirmohamed, M. (2013). Clinical Pharmacogenetics Implementation Consortium guidelines for HLA‐B genotype and carbamazepine dosing. Clinical Pharmacology & Therapeutics *94*, 324-328.
Leslie, R., O'Donnell, C.J., and Johnson, A.D. (2014). GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics *30*, i185-i194.
Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research *44*, W242-W245.
Li, A., Dubey, S., Varney, M.L., Dave, B.J., and Singh, R.K. (2003). IL-8 directly enhanced endothelial cell survival, proliferation, and matrix metalloproteinases production and regulated angiogenesis. The Journal of Immunology *170*, 3369-3376.
Li, A., Dubey, S., Varney, M.L., and Singh, R.K. (2002). Interleukin-8-induced proliferation, survival, and MMP production in CXCR1 and CXCR2 expressing human umbilical vein endothelial cells. Microvascular research *64*, 476-481.
Li, A., Varney, M.L., and Singh, R.K. (2001). Expression of interleukin 8 and its receptors in human colon carcinoma cells with different metastatic potentials. Clinical Cancer Research *7*, 3298-3304.
Li, A., Varney, M.L., Valasek, J., Godfrey, M., Dave, B.J., and Singh, R.K. (2005). Autocrine role of interleukin-8 in induction of endothelial cell proliferation, survival, migration and MMP-2 production and angiogenesis. Angiogenesis *8*, 63-71.

Li, H., Tian, J., Cheng, H., Zhang, H., and Zhang, D. (2016). Association of tumor necrosis factor alpha polymorphisms with cervical cancer in a Chinese population. INTERNATIONAL JOURNAL OF CLINICAL AND EXPERIMENTAL PATHOLOGY *9*, 2200-2207.
Li, H., Yi, T., Zhao, S., Chen, P., Cheng, C., Wei, Y., and Zhao, X. (2009). The anti‐condyloma acuminatum effects of interferon‐inducible protein 10 in vitro. International journal of dermatology *48*, 136-141.
Li, M.-Q., Luo, X.-Z., Meng, Y.-H., Mei, J., Zhu, X.-Y., Jin, L.-P., and Li, D.-J. (2012). CXCL8 enhances proliferation and growth and reduces apoptosis in endometrial stromal cells in an autocrine manner via a CXCR1-triggered PTEN/AKT signal pathway. Human reproduction *27*, 2107-2116.
Li, M., Han, Y., Wu, T.-T., Feng, Y., and Wang, H.-B. (2013). Tumor necrosis factor alpha rs1800629 polymorphism and risk of cervical lesions: a meta-analysis. PloS one *8*, e69201.
Li, Q., Cheng, X., Ji, J., Zhang, J., and Zhou, X. (2014). Gene amplification of EGFR and its clinical significance in various cervical (lesions) lesions using cytology and FISH. International journal of clinical and experimental pathology *7*, 2477.
Li, X., Huang, K., Zhang, Q., Zhou, J., Sun, H., Tang, F., Zhou, H., Hu, T., Wang, S., and Jia, Y. (2017). Genome-wide association study identifies four SNPs associated with response to platinum-based neoadjuvant chemotherapy for cervical cancer. Scientific reports *7*.
Lima, M., Leander, M., Santos, M., Santos, A.H., Lau, C., Queirós, M.L., Gonçalves, M., Fonseca, S., Moura, J., and Teixeira, M.d.A. (2015). Chemokine Receptor Expression on Normal Blood CD56. Journal of immunology research *2015*.
Liu, F., Lin, B., Liu, X., Zhang, W., Zhang, E., Hu, L., Ma, Y., Li, X., and Tang, X. (2016a). ERK Signaling Pathway Is Involved in HPV-16 E6 but not E7 Oncoprotein-Induced HIF-1α Protein Accumulation in NSCLC Cells. Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics *23*, 109-118.
Liu, L.-B., Xie, F., Chang, K.-K., Li, M.-Q., Meng, Y.-H., Wang, X.-H., Li, H., Li, D.-J., and Yu, J.-J. (2014). Hypoxia promotes the proliferation of cervical carcinoma cells through stimulating the secretion of IL-8. International journal of clinical and experimental pathology *7*, 575.
Liu, L., Yang, X., Chen, X., Kan, T., Shen, Y., Chen, Z., and Hu, Z. (2012a). Association between TNF-alpha polymorphisms and cervical cancer risk: a meta-analysis. Molecular Biology Reports *39*, 2683-2688.
Liu, Q., Li, A., Tian, Y., Wu, J.D., Liu, Y., Li, T., Chen, Y., Han, X., and Wu, K. (2016b). The CXCL8-CXCR1/2 pathways in cancer. Cytokine & growth factor reviews *31*, 61-71.
Liu, X., Peng, J., Sun, W., Yang, S., Deng, G., Li, F., Cheng, J.-W., and Gordon, J.R. (2012b). G31P, an antagonist against CXC chemokine receptors 1 and 2, inhibits growth of human prostate cancer cells in nude mice. The Tohoku journal of experimental medicine *228*, 147-156.
Liu, Y., Yang, S., Lin, A.A., Cavalli-Sforza, L.L., and Su, B. (2005). Molecular evolution of CXCR1, a G protein-coupled receptor involved in signal transduction of neutrophils. Journal of molecular evolution *61*, 691-696.
Lowman, H.B., Slagle, P.H., DeForge, L.E., Wirth, C.M., Gillece-Castro, B.L., Bourell, J.H., and Fairbrother, W.J. (1996). Exchanging interleukin-8 and melanoma growth-stimulating activity receptor binding specificities. Journal of Biological Chemistry *271*, 14344-14352.
Lundstedt, A.-C., McCarthy, S., Gustafsson, M.C., Godaly, G., Jodal, U., Karpman, D., Leijonhufvud, I., Lindén, C., Martinell, J., and Ragnarsdottir, B. (2007). A genetic basis of susceptibility to acute pyelonephritis. PLoS One *2*, e825.
Lunetta, K.L. (2008). Genetic association studies. Circulation *118*, 96-101.
Luo, Y., de Lange, K.M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N.A., Lamb, C.A., McCarthy, S., Ahmad, T., and Edwards, C. (2017). Exploring the genetic architecture of

inflammatory bowel disease by whole genome sequencing identifies association at ADCY7. Nature genetics *49*, 186.
Lutkowska, A., Roszak, A., Lianeri, M., Sowińska, A., Sotiri, E., and Jagodziński, P.P. (2017). Analysis of rs8067378 Polymorphism in the Risk of Uterine Cervical Cancer from a Polish Population and its Impact on Gasdermin B Expression. Molecular diagnosis & therapy *21*, 199-207.
Müller, M., Prescott, E.L., Wasson, C.W., and Macdonald, A. (2015). Human papillomavirus E5 oncoprotein: function and potential target for antiviral therapeutics. Future Virology *10*, 27-39.
MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., and Morales, J. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research, gkw1133.
Magnusson, P.K., Lichtenstein, P., and Gyllensten, U.B. (2000). Heritability of cervical tumours. International journal of cancer *88*, 698-701.
Maley, S.N., Schwartz, S.M., Johnson, L.G., Malkki, M., Du, Q., Daling, J.R., Li, S.S., Zhao, L.P., Petersdorf, E.W., and Madeleine, M.M. (2009). Genetic variation in CXCL12 and risk of cervical carcinoma: a population‐based case–control study. International journal of immunogenetics *36*, 367-375.
Mamik, M.K., and Ghorpade, A. (2014). Chemokine CXCL8 promotes HIV-1 replication in human monocyte-derived macrophages and primary microglia via nuclear factor-κB pathway. PloS one *9*, e92145.
Martínez-Nava, G.A., Fernández-Niño, J.A., Madrid-Marina, V., and Torres-Poveda, K. (2016). Cervical cancer genetic susceptibility: A systematic review and meta-analyses of recent evidence. PloS one *11*, e0157344.
Marth, C., Landoni, F., Mahner, S., McCormack, M., Gonzalez-Martin, A., Colombo, N., and Committee, E.G. (2017). Cervical cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology *28*, iv72-iv83.
Martin, M., Klein, T., Dong, B., Pirmohamed, M., Haas, D., and Kroetz, D. (2012). Clinical pharmacogenetics implementation consortium guidelines for HLA‐B genotype and abacavir dosing. Clinical Pharmacology & Therapeutics *91*, 734-738.
Marullo, R., Werner, E., Zhang, H., Chen, G.Z., Shin, D.M., and Doetsch, P.W. (2015). HPV16 E6 and E7 proteins induce a chronic oxidative stress response via NOX2 that causes genomic instability and increased susceptibility to DNA damage in head and neck cancer cells. Carcinogenesis *36*, 1397-1406.
McBride, A.A. (2013). The papillomavirus E2 proteins. Virology *445*, 57-79.
McQuillan, G., Kruszon-Moran, D., Markowitz, L., Unger, E., and Paulose-Ram, R. (2017). Prevalence of HPV in Adults Aged 18-69: United States, 2011-2014. NCHS data brief, 1.
Mehta, A.M., Mooij, M., Branković, I., Ouburg, S., Morré, S.A., and Jordanova, E.S. (2017). Cervical Carcinogenesis and Immune Response Gene Polymorphisms: A Review. Journal of immunology research *2017*.
Mehta, A.M., Spaans, V.M., Mahendra, N.B., Osse, E.M., Vet, J.N., Purwoto, G., Surya, I.G., Cornian, S., Peters, A.A., Fleuren, G.J.*, et al.* (2015). Differences in genetic variation in antigen-processing machinery components and association with cervical carcinoma risk in two Indonesian populations. Immunogenetics *67*, 267-275.
Mitt, M., Kals, M., Parn, K., Gabriel, S.B., Lander, E.S., Palotie, A., Ripatti, S., Morris, A.P., Metspalu, A., and Esko, T. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. European Journal of Human Genetics.

Miura, K., Mishima, H., Kinoshita, A., Hayashida, C., Abe, S., Tokunaga, K., Masuzaki, H., and Yoshiura, K. (2014). Genome-wide association study of HPV-associated cervical cancer in Japanese women. Journal of Medical Virology *86*, 1153-1158.
Miura, K., Mishima, H., Yasunami, M., Kaneuchi, M., Kitajima, M., Abe, S., Higashijima, A., Fuchi, N., Miura, S., and Yoshiura, K.-I. (2016). A significant association between rs8067378 at 17q12 and invasive cervical cancer originally identified by a genome-wide association study in Han Chinese is replicated in a Japanese population. Journal of human genetics *61*, 793-796.
Moepps, B., Nuesseler, E., Braun, M., and Gierschik, P. (2006). A homolog of the human chemokine receptor CXCR1 is expressed in the mouse. Molecular immunology *43*, 897-914.
Moody, C.A., and Laimins, L.A. (2010). Human papillomavirus oncoproteins: pathways to transformation. Nature reviews Cancer *10*, 550-560.
Moscicki, A.-B., Ma, Y., Wibbelsman, C., Darragh, T.M., Powers, A., Farhat, S., and Shiboski, S. (2010). Rate of and risks for regression of CIN-2 in adolescents and young women. Obstetrics and gynecology *116*, 1373.
Motamedi, M., Böhmer, G., Neumann, H.H., and von Wasielewski, R. (2015). CIN III lesions and regression: retrospective analysis of 635 cases. BMC infectious diseases *15*, 541.
Murphy, C., McGurk, M., Pettigrew, J., Santinelli, A., Mazzucchelli, R., Johnston, P.G., Montironi, R., and Waugh, D.J. (2005). Nonapical and cytoplasmic expression of interleukin-8, CXCR1, and CXCR2 correlates with cell proliferation and microvessel density in prostate cancer. Clinical Cancer Research *11*, 4117-4127.
Nagarsheth, N., Wicha, M.S., and Zou, W. (2017). Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. Nature Reviews Immunology.
Namipashaki, A., Razaghi-Moghadam, Z., and Ansari-Pour, N. (2015). The essentiality of reporting Hardy-Weinberg equilibrium calculations in population-based genetic association studies. Cell Journal (Yakhteh) *17*, 187.
Nasser, M.W., Raghuwanshi, S.K., Malloy, K.M., Gangavarapu, P., Shim, J.-Y., Rajarathnam, K., and Richardson, R.M. (2007). CXCR1 and CXCR2 Activation and Regulation ROLE OF ASPARTATE 199 OF THE SECOND EXTRACELLULAR LOOP OF CXCR2 IN CXCL8-MEDIATED RAPID RECEPTOR INTERNALIZATION. Journal of Biological Chemistry *282*, 6906-6915.
Network, C.G.A.R. (2017). Integrated genomic and molecular characterization of cervical cancer. Nature *543*, 378.
Nieves-Ramirez, M.E., Partida-Rodriguez, O., Alegre-Crespo, P.E., del Carmen Tapia-Lugo, M., and Perez-Rodriguez, M.E. (2011). Characterization of single-nucleotide polymorphisms in the tumor necrosis factor α promoter region and in lymphotoxin α in squamous intraepithelial lesions, precursors of cervical cancer. Translational oncology *4*, 336-344.
Okada, Y., Meguro, M., Ohyama, H., Yoshizawa, S., Takeuchi‐Hatanaka, K., Kato, N., Matsushita, S., Takashiba, S., and Nishimura, F. (2009). Human leukocyte histocompatibility antigen class II‐induced cytokines from human gingival fibroblasts promote proliferation of human umbilical vein endothelial cells: potential association with enhanced angiogenesis in chronic periodontal inflammation. Journal of periodontal research *44*, 103-109.
Okolicsanyi, R.K., Faure, M., Jacinto, J.M., Chacon-Cortes, D., Chambers, S., Youl, P.H., Haupt, L.M., and Griffiths, L.R. (2014). Association of the SNP rs2623047 in the HSPG modification enzyme SULF1 with an Australian Caucasian Breast Cancer Cohort. Gene *547*, 50-54.
Paavonen, J., Naud, P., Salmerón, J., Wheeler, C.M., Chow, S.-N., Apter, D., Kitchener, H., Castellsagué, X., Teixeira, J.C., and Skinner, S.R. (2009). Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infection and precancer caused by

oncogenic HPV types (PATRICIA): final analysis of a double-blind, randomised study in young women. The Lancet *374*, 301-314.
Palter, S.F., Mulayim, N., Senturk, L., and Arici, A. (2001). Interleukin-8 in the human fallopian tube. The Journal of Clinical Endocrinology & Metabolism *86*, 2660-2667.
Pananghat, A.N., Aggarwal, H., Prakash, S.S., Makhdoomi, M.A., Singh, R., Lodha, R., Ali, S., Srinivas, M., Das, B.K., and Pandey, R.M. (2016). IL-8 Alterations in HIV-1 Infected Children With Disease Progression. Medicine *95*.
Paneth, N., Susser, E., and Susser, M. (2002). Origins and early development of the case-control study: Part 1, Early evolution. Sozial-und Präventivmedizin/Social and Preventive Medicine *47*, 282-288.
Paradisi, A., Waterboer, T., Sampogna, F., Tabolli, S., Simoni, S., Pawlita, M., and Abeni, D. (2011). Seropositivity for human papillomavirus and incidence of subsequent squamous cell and basal cell carcinomas of the skin in patients with a previous nonmelanoma skin cancer. British Journal of Dermatology *165*, 782-791.
Paradkar, P.H., Joshi, J.V., Mertia, P.N., Agashe, S.V., and Vaidya, R.A. (2014). Role of cytokines in genesis, progression and prognosis of cervical cancer. Asian Pac J Cancer Prev *15*, 3851-3864.
Park, M.S., Kim, Y.H., Jung, Y., Kim, S.H., Park, J.C., Yoon, D.S., Kim, S.-H., and Lee, J.W. (2015). In situ recruitment of human bone marrow-derived mesenchymal stem cells using chemokines for articular cartilage regeneration. Cell transplantation *24*, 1067-1083.
Park, S.H., Das, B.B., Casagrande, F., Tian, Y., Nothnagel, H.J., Chu, M., Kiefer, H., Maier, K., De Angelis, A., and Marassi, F.M. (2012). Structure of the chemokine receptor CXCR1 in phospholipid bilayers. Nature *491*, 779.
Pascale, R.M., Calvisi, D.F., and Feo, F. (2016). Sulfatase 1: a new Jekyll and Hyde in hepatocellular carcinoma? Translational Gastroenterology and Hepatology *1*.
Pecorelli, S. (2009). Revised FIGO staging for carcinoma of the vulva, cervix, and endometrium. International Journal of Gynecology & Obstetrics *105*, 103-104.
Peitsaro, P., Johansson, B., and Syrjänen, S. (2002). Integrated human papillomavirus type 16 is frequently found in cervical cancer precursors as demonstrated by a novel quantitative real-time PCR technique. Journal of clinical microbiology *40*, 886-891.
Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nature genetics.
Raff, A.B., Woodham, A.W., Raff, L.M., Skeate, J.G., Yan, L., Da Silva, D.M., Schelhaas, M., and Kast, W.M. (2013). The evolving field of human papillomavirus receptor research: a review of binding and entry. Journal of virology *87*, 6062-6072.
Ragnarsdóttir, B., Lutay, N., Grönberg-Hernandez, J., Köves, B., and Svanborg, C. (2011). Genetics of innate immunity and UTI susceptibility. Nature Reviews Urology *8*, 449-468.
Ragnarsdottir, B., Fischer, H., Godaly, G., Grönberg‐Hernandez, J., Gustafsson, M., Karpman, D., Lundstedt, A.-C., Lutay, N., Rämisch, S., and Svensson, M. (2008). TLR‐and CXCR1‐dependent innate immunity: insights into the genetics of urinary tract infections. European journal of clinical investigation *38*, 12-20.
Rajagopalan, L., and Rajarathnam, K. (2004). Ligand selectivity and affinity of chemokine receptor CXCR1 Role of N-terminal domain. Journal of Biological Chemistry *279*, 30000-30008.
Raychaudhuri, B., and Vogelbaum, M.A. (2011). IL-8 is a mediator of NF-κB induced invasion by gliomas. Journal of neuro-oncology *101*, 227-235.
Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., and Ward, R. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199-204.

Reich, O., Regauer, S., Marth, C., Schmidt, D., Horn, L.-C., Dannecker, C., Menton, M., and Beckmann, M. (2015). Precancerous lesions of the cervix, vulva and vagina according to the 2014 WHO classification of tumors of the female genital tract. Geburtshilfe und Frauenheilkunde *75*, 1018-1020.
Relling, M., Gardner, E., Sandborn, W., Schmiegelow, K., Pui, C.H., Yee, S., Stein, C., Carrillo, M., Evans, W., and Klein, T. (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clinical Pharmacology & Therapeutics *89*, 387-391.
Reyes-Robles, T., Alonzo, F., Kozhaya, L., Lacy, D.B., Unutmaz, D., and Torres, V.J. (2013). Staphylococcus aureus leukotoxin ED targets the chemokine receptors CXCR1 and CXCR2 to kill leukocytes and promote infection. Cell host & microbe *14*, 453-459.
Richardson, R.M., Marjoram, R.J., Barak, L.S., and Snyderman, R. (2003). Role of the cytoplasmic tails of CXCR1 and CXCR2 in mediating leukocyte migration, activation, and regulation. The Journal of Immunology *170*, 2904-2911.
Ringe, J., Strassburg, S., Neumann, K., Endres, M., Notter, M., Burmester, G.R., Kaps, C., and Sittinger, M. (2007). Towards in situ tissue repair: human mesenchymal stem cells express chemokine receptors CXCR1, CXCR2 and CCR2, and migrate upon stimulation with CXCL8 but not CCL2. Journal of cellular biochemistry *101*, 135-146.
Rotar, I.C., MUREȘAN, D., Radu, P., PETRIȘOR, F., Apostol, S., Mariana, T., Butuza, C., and Stamatian, F. (2014). TNF-α 308 G/A polymorphism and cervical intraepithelial neoplasia. Anticancer research *34*, 373-378.
Russo, R.C., Garcia, C.C., Teixeira, M.M., and Amaral, F.A. (2014). The CXCL8/IL-8 chemokine family and its receptors in inflammatory diseases. Expert review of clinical immunology *10*, 593-619.
Safaeian, M., Hildesheim, A., Gonzalez, P., Yu, K., Porras, C., Li, Q., Rodriguez, A.C., Sherman, M.E., Schiffman, M., Wacholder, S.*, et al.* (2012). Single nucleotide polymorphisms in the PRDX3 and RPS19 and risk of HPV persistence and cervical precancer/cancer. PLoS ONE [Electronic Resource] *7*, e33619.
Sales, K.J., Sutherland, J.R., Jabbour, H.N., and Katz, A.A. (2012). Seminal plasma induces angiogenic chemokine expression in cervical cancer cells and regulates vascular function. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research *1823*, 1789-1795.
Salter, J., and Demos (2014). Behind the Screen: Revealing the True Cost of Cervical Cancer (Demos).
Sato, E., Fujimoto, J., Toyoki, H., Sakaguchi, H., Alam, S., Jahan, I., and Tamaya, T. (2007). Expression of IP-10 related to angiogenesis in uterine cervical cancers. British journal of cancer *96*, 1735-1739.
Schierding, W., Cutfield, W.S., and O'Sullivan, J.M. (2014). The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell. Frontiers in genetics *5*.
Schmausser, B., Josenhans, C., Endrich, S., Suerbaum, S., Sitaru, C., Andrulis, M., Brändlein, S., Rieckmann, P., Müller-Hermelink, H.K., and Eck, M. (2004). Downregulation of CXCR1 and CXCR2 Expression on Human Neutrophils by Helicobacter pylori: a New Pathomechanism in H. pylori Infection? Infection and immunity *72*, 6773-6779.
Schraufstatter, I.U., Chung, J., and Burger, M. (2001). IL-8 activates endothelial cell CXCR1 and CXCR2 through Rho and Rac signaling pathways. American Journal of Physiology-Lung Cellular and Molecular Physiology *280*, L1094-L1103.
Schroer, N., Pahne, J., Walch, B., Wickenhauser, C., and Smola, S. (2011). Molecular pathobiology of human cervical high-grade lesions: paracrine STAT3 activation in tumor-instructed myeloid cells drives local MMP-9 expression. Cancer research *71*, 87-97.

Sfakianakis, A., Barr, C.E., and Kreutzer, D.L. (2002). Localization of the chemokine interleukin‐8 and interleukin‐8 receptors in human gingiva and cultured gingival keratinocytes. Journal of periodontal research *37*, 154-160.
Shamaladevi, N., Lyn, D.A., Escudero, D.O., and Lokeshwar, B.L. (2009). CXC receptor-1 silencing inhibits androgen-independent prostate cancer. Cancer research *69*, 8265-8274.
Sheet, C.F. (2007). Genital HPV (Dec).
Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research *29*, 308-311.
Shi, Y., Li, L., Hu, Z., Li, S., Wang, S., Liu, J., Wu, C., He, L., Zhou, J., Li, Z.*, et al.* (2013). A genome-wide association study identifies two new cervical cancer susceptibility loci at 4q12 and 17q12. Nature Genetics *45*, 918-922.
Singh, H., Sachan, R., Jain, M., and Mittal, B. (2008). CCR5-Δ32 Polymorphism and Susceptibility to Cervical Cancer: Association With Early Stage of Cervical Cancer. Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics *17*, 87-91.
Singh, S., Sadanandam, A., Varney, M.L., Nannuru, K.C., and Singh, R.K. (2010). Small interfering RNA‐mediated CXCR1 or CXCR2 knock‐down inhibits melanoma tumor growth and invasion. International journal of cancer *126*, 328-336.
Singh, S., Wu, S., Varney, M., Singh, A.P., and Singh, R.K. (2011). CXCR1 and CXCR2 silencing modulates CXCL8-dependent endothelial cell proliferation, migration and capillary-like structure formation. Microvascular research *82*, 318-325.
Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J.G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J.F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. The American Journal of Human Genetics *89*, 607-618.
Solé, X., Guinó, E., Valls, J., Iniesta, R., and Moreno, V. (2006). SNPStats: a web tool for the analysis of association studies. Bioinformatics *22*, 1928-1929.
Soto, D., Song, C., and McLaughlin-Drubin, M.E. (2017). Epigenetic Alterations in Human Papillomavirus-Associated Cancers. Viruses *9*.
Sousa, H., Oliveira, S., Santos, A.M., Catarino, R., Moutinho, J., and Medeiros, R. (2014). Tumour necrosis factor alpha 308 G/A is a risk marker for the progression from high-grade lesions to invasive cervical cancer. Tumor Biology *35*, 2561-2564.
Spathis, A., Aga, E., Alepaki, M., Chranioti, A., Meristoudis, C., Panayiotides, I., Kassanos, D., and Karakitsos, P. (2011). Promoter methylation of p16INK4A, hMLH1, and MGMT in liquid-based cervical cytology samples compared with clinicopathological findings and HPV presence. Infectious diseases in obstetrics and gynecology *2011*.
Srivenugopal, K.S., and Ali-Osman, F. (2002). The DNA repair protein, O(6)-methylguanine-DNA methyltransferase is a proteolytic target for the E6 human papillomavirus oncoprotein. Oncogene *21*, 5940-5945.
Stearns, F.W. (2010). One hundred years of pleiotropy: a retrospective. Genetics *186*, 767-773.
Stillie, R., Farooq, S.M., Gordon, J.R., and Stadnyk, A.W. (2009). The functional significance behind expressing two IL-8 receptor types on PMN. Journal of leukocyte biology *86*, 529-543.
Stram, D.O. (2014). Design of large-scale genetic association studies, sample size, and power. In Design, Analysis, and Interpretation of Genome-Wide Association Scans (Springer), pp. 243-284.
Sun, L.-l., Cao, D.-y., Yang, J.-x., Li, H., Zhou, X.-r., Song, Z.-q., Cheng, X.-m., Chen, J., and Shen, K. (2012). Population-based case–control study on DAPK1, RAR-β2 and MGMT methylation in liquid-based cytology. Archives of gynecology and obstetrics *285*, 1433-1439.

Swamydas, M., Gao, J.-L., Break, T.J., Johnson, M.D., Jaeger, M., Rodriguez, C.A., Lim, J.K., Green, N.M., Collar, A.L., and Fischer, B.G. (2016). CXCR1-mediated neutrophil degranulation and fungal killing promote Candida clearance and host survival. Science translational medicine *8*, 322ra310-322ra310.
Tabata, S., Ikeda, R., Yamamoto, M., Furukawa, T., Kuramoto, T., Takeda, Y., Yamada, K., Haraguchi, M., Nishioka, Y., and Sone, S. (2012). Thymidine phosphorylase enhances reactive oxygen species generation and interleukin-8 expression in human cancer cells. Oncology reports *28*, 895-902.
Takata, H., Tomiyama, H., Fujiwara, M., Kobayashi, N., and Takiguchi, M. (2004). Cutting edge: expression of chemokine receptor CXCR1 on human effector CD8+ T cells. The Journal of Immunology *173*, 2231-2235.
Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Research.
Tjiong, M.Y., van der Vange, N., ten Kate, F.J., Tjong-A-Hung, S.P., ter Schegget, J., Burger, M.P., and Out, T.A. (1999). Increased IL-6 and IL-8 levels in cervicovaginal secretions of patients with cervical cancer. Gynecologic oncology *73*, 285-291.
Tommasino, M. (2014). The human papillomavirus family and its role in carcinogenesis. Paper presented at: Seminars in cancer biology (Elsevier).
Tseng-Rogenski, S., and Liebert, M. (2009). Interleukin-8 is essential for normal urothelial cell survival. American Journal of Physiology-Renal Physiology *297*, F816-F821.
Uen, W.-C., Hsieh, C.-H., Tseng, T.-T., Jiang, S.S., Tseng, J.-C., and Lee, S.-C. (2015). Anchorage independency promoted tumor malignancy of melanoma cells under reattachment through elevated interleukin-8 and CXC chemokine receptor 1 expression. Melanoma research *25*, 35-46.
Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S., Gopalan, V., Mohamoud, Y., Huyen, Y., and McBride, A.A. (2013). The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic acids research *41*, D571-D578.
Vieira, R., Simões, M.J., Carmona, S., Egas, C., Faro, C., and Figueiredo, A. (2013). Identification of DLEC1 D215N Somatic Mutation in Formalin Fixed Paraffin Embedded Melanoma and Melanocytic Nevi Specimens. Journal of skin cancer *2013*.
Vinet, J., Zwam, M., Dijkstra, I., Brouwer, N., Weering, H., Watts, A., Meijer, M., Fokkens, M., Kannan, V., and Verzijl, D. (2013). Inhibition of CXCR3‐mediated chemotaxis by the human chemokine receptor‐like protein CCX‐CKR. British journal of pharmacology *168*, 1375-1387.
Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics *101*, 5-22.
Walker, J., Smiley, L.C., Ingram, D., and Roman, A. (2011). Expression of human papillomavirus type 16 E7 is sufficient to significantly increase expression of angiogenic factors but is not sufficient to induce endothelial cell migration. Virology *410*, 283-290.
Wallace, N.A., Khanal, S., Robinson, K.L., Wendel, S.O., Messer, J.J., and Galloway, D.A. (2017). High-Risk Alphapapillomavirus Oncogenes Impair the Homologous Recombination Pathway. J Virol *91*.
Wang, J., Hu, W., Wu, X., Wang, K., Yu, J., Luo, B., Luo, G., Wang, W., Wang, H., and Li, J. (2016a). CXCR1 promotes malignant behavior of gastric cancer cells in vitro and in vivo in AKT and ERK1/2 phosphorylation. International journal of oncology *48*, 2184-2196.
Wang, S., Sun, H., Jia, Y., Tang, F., Zhou, H., Li, X., Zhou, J., Huang, K., Zhang, Q., and Hu, T. (2015). Association of 42 SNPs with genetic risk for cervical cancer: an extensive meta-analysis. BMC medical genetics *16*, 25.

Wang, S.S., Bratti, M.C., Rodríguez, A.C., Herrero, R., Burk, R.D., Porras, C., González, P., Sherman, M.E., Wacholder, S., and Lan, Z.E. (2009a). Common variants in immune and DNA repair genes and risk for human papillomavirus persistence and progression to cervical cancer. Journal of Infectious Diseases *199*, 20-30.
Wang, S.S., Gonzalez, P., Yu, K., Porras, C., Li, Q., Safaeian, M., Rodriguez, A.C., Sherman, M.E., Bratti, C., Schiffman, M*., et al.* (2010). Common genetic variants and risk for HPV persistence and progression to cervical cancer. PLoS ONE [Electronic Resource] *5*, e8667.
Wang, S.S., Zuna, R.E., Wentzensen, N., Dunn, S.T., Sherman, M.E., Gold, M.A., Schiffman, M., Wacholder, S., Allen, R.A., and Block, I. (2009b). Human papillomavirus cofactors by disease progression and human papillomavirus types in the study to understand cervical cancer early endpoints and determinants. Cancer Epidemiology and Prevention Biomarkers *18*, 113-120.
Wang, X., Li, T., Wang, W., Yuan, W., Liu, H., Cheng, Y., Wang, P., Zhang, Y., and Han, W. (2016b). Cytokine-like 1 chemoattracts monocytes/macrophages via CCR2. The Journal of Immunology *196*, 4090-4099.
Ward, L.D., and Kellis, M. (2011). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research *40*, D930-D934.
Waugh, D.J., and Wilson, C. (2008). The interleukin-8 pathway in cancer. Clinical cancer research *14*, 6735-6741.
Waxman, A.G., Chelmow, D., Darragh, T.M., Lawson, H., and Moscicki, A.-B. (2012). Revised terminology for cervical histopathology and its implications for management of high-grade squamous intraepithelial lesions of the cervix. Obstetrics and gynecology *120*, 1465.
Wei, J., Peng, J., Wang, B., Qu, H., Wang, S., Faisal, A., Cheng, J.-W., Gordon, J.R., and Li, F. (2013). CXCR1/CXCR2 antagonism is effective in pulmonary defense against Klebsiella pneumoniae infection. BioMed research international *2013*.
Wei, M., Liang, L.Z., Zhang, C.Q., Xiong, Y., Zhang, Y., Shen, Y., and Li, J.Q. (2007). [Correlation of CXCR4/CXCL12 overexpression to lymph node metastasis and chronic inflammation in cervical adenocarcinoma]. Ai zheng = Aizheng = Chinese journal of cancer *26*, 298-302.
Whittall, C., Kehoe, O., King, S., Rot, A., Patterson, A., and Middleton, J. (2013). A chemokine self-presentation mechanism involving formation of endothelial surface microstructures. The Journal of Immunology *190*, 1725-1736.
Wigginton, J.E., Cutler, D.J., and Abecasis, G.R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. The American Journal of Human Genetics *76*, 887-893.
Wilson, B.J., Saab, K.R., Ma, J., Schatton, T., Pütz, P., Zhan, Q., Murphy, G.F., Gasser, M., Waaga-Gasser, A.M., and Frank, N.Y. (2014). ABCB5 maintains melanoma-initiating cells through a proinflammatory cytokine signaling circuit. Cancer research *74*, 4196-4207.
Wilting, S.M., and Steenbergen, R.D.M. (2016). Molecular events leading to HPV-induced high grade neoplasia. Papillomavirus research (Amsterdam, Netherlands) *2*, 85-88.
Wong, K.M., Langlais, K., Tobias, G.S., Fletcher-Hoppe, C., Krasnewich, D., Leeds, H.S., Rodriguez, L.L., Godynskiy, G., Schneider, V.A., and Ramos, E.M. (2017). The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. Nucleic Acids Research *45*, D819-D826.
Wu, H.-H., Lee, T.-H., Tee, Y.-T., Chen, S.-C., Yang, S.-F., Lee, S.-K., Ko, J.-L., and Wang, P.-H. (2013a). Relationships of single nucleotide polymorphisms of monocyte chemoattractant protein 1 and chemokine receptor 2 with susceptibility and clinicopathologic characteristics of neoplasia of uterine cervix in Taiwan women. Reproductive Sciences *20*, 1175-1183.
Wu, S., Lu, S., Tao, H., Zhang, L., Lin, W., Shang, H., and Xie, J. (2011). Correlation of polymorphism of IL-8 and MMP-7 with occurrence and lymph node metastasis of early

stage cervical cancer. Journal of Huazhong University of Science and Technology [Medical Sciences] *31*, 114-119.
Wu, S., Shang, H., Cui, L., Zhang, Z., Zhang, Y., Li, Y., Wu, J., Li, R.-K., and Xie, J. (2013b). Targeted blockade of interleukin-8 abrogates its promotion of cervical cancer growth and metastasis. Molecular and cellular biochemistry *375*, 69-79.
Wuyts, A., Proost, P., Lenaerts, J.P., Ben‐Baruch, A., Van Damme, J., and Wang, J.M. (1998). Differential usage of the CXC chemokine receptors 1 and 2 by interleukin‐8, granulocyte chemotactic protein‐2 and epithelial‐cell‐derived neutrophil attractant‐78. The FEBS Journal *255*, 67-73.
Xue, J., Vesper, B.J., and Radosevich, J.A. (2012a). Human Papillomavirus: A Brief Overview. In HPV and Cancer (Springer), pp. 1-15.
Xue, J., Vesper, B.J., and Radosevich, J.A. (2012b). The life cycle of human papillomavirus. In HPV and Cancer (Springer), pp. 49-74.
Xue, J., Vesper, B.J., and Radosevich, J.A. (2012c). Proteins Encoded by the Human Papillomavirus Genome and Their Functions. In HPV and Cancer (Springer), pp. 17-47.
Yang, J.D., Sun, Z., Hu, C., Lai, J., Dove, R., Nakamura, I., Lee, J.S., Thorgeirsson, S.S., Kang, K.J., and Chu, I.S. (2011). Sulfatase 1 and sulfatase 2 in hepatocellular carcinoma: associated signaling pathways, tumor phenotypes, and survival. Genes, Chromosomes and Cancer *50*, 122-135.
Yang, L., Ping, C., Luo, S., Li, J., Liu, K., Hu, H., and Wei, Y. (2009). CXCL10 gene therapy efficiently inhibited the growth of cervical carcinoma based on the antiangiogenic and antiviral activity. Biotechnology and applied biochemistry.
Yoshimura, T., Matsushima, K., Tanaka, S., Robinson, E.A., Appella, E., Oppenheim, J.J., and Leonard, E.J. (1987). Purification of a human monocyte-derived neutrophil chemotactic factor that has peptide sequence similarity to other host defense cytokines. Proceedings of the National Academy of Sciences *84*, 9233-9237.
Yu, K., Li, Q., Bergen, A.W., Pfeiffer, R.M., Rosenberg, P.S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P‐values. Genetic epidemiology *33*, 700-709.
Zeng, B., Lloyd-Jones, L.R., Holloway, A., Marigorta, U.M., Metspalu, A., Montgomery, G.W., Esko, T., Brigham, K.L., Quyyumi, A.A., and Idaghdour, Y. (2017). Constraints on eQTL fine mapping in the presence of multi-site local regulation of gene expression. G3: Genes, Genomes, Genetics, g3. 117.043752.
Zhang, T., Tseng, C., Zhang, Y., Sirin, O., Corn, P.G., Li-Ning-Tapia, E.M., Troncoso, P., Davis, J., Pettaway, C., and Ward, J. (2016). CXCL1 mediates obesity-associated adipose stromal cell trafficking and function in the tumour microenvironment. Nature communications *7*, 11674.
Zhang, W., Liu, Y., Zhao, N., Chen, H., Qiao, L., Zhao, W., and Chen, J.J. (2015). Role of Cdk1 in the p53-independent abrogation of the postmitotic checkpoint by human papillomavirus E6. J Virol *89*, 2553-2562.
Zhang, Y., Wu, J.Z., Yang, Y.Q., Ma, R., Zhang, J.Y., and Feng, J.F. (2014). Expression of growth‐regulated oncogene‐1, hepatocyte growth factor, platelet‐derived growth factor‐AA and soluble E‐selectin and their association with high‐risk human papillomavirus infection in squamous cell carcinoma of the uterine cervix. Molecular medicine reports *10*, 1013-1024.
Zhang, Z., Borecki, I., Nguyen, L., Ma, D., Smith, K., Huettner, P.C., Mutch, D.G., Herzog, T.J., Gibb, R.K., Powell, M.A*., et al.* (2007). CD83 gene polymorphisms increase susceptibility to human invasive cervical cancer. Cancer research *67*, 11202-11208.

Zhao, M., Qiu, L., Tao, N., Zhang, L., Wu, X., She, Q., Zeng, F., Wang, Y., Wei, S., and Wu, X. (2013). HLA DRB allele polymorphisms and risk of cervical cancer associated with human papillomavirus infection: a population study in China. European Journal of Gynaecological Oncology *34*, 54-59.
Zheng, B., Wiklund, F., Gharizadeh, B., Sadat, M., Gambelunghe, G., Hallmans, G., Dillner, J., Wallin, K.-L., and Ghaderi, M. (2006). Genetic polymorphism of chemokine receptors CCR2 and CCR5 in Swedish cervical cancer patients. Anticancer research *26*, 3669-3674.
Zhou, J., Yi, L., Ouyang, Q., Xu, L., Cui, H., and Xu, M. (2014). Neurotensin signaling regulates stem-like traits of glioblastoma stem cells through activation of IL-8/CXCR1/STAT3 pathway. Cellular signalling *26*, 2896-2902.
Zhu, M., and Zhao, S. (2007). Candidate gene identification approach: progress and challenges. International journal of biological sciences *3*, 420.
Zijlmans, H.J., Fleuren, G.J., Baelde, H.J., Eilers, P.H., Kenter, G.G., and Gorter, A. (2006). The absence of CCL2 expression in cervical carcinoma is associated with increased survival and loss of heterozygosity at 17q11.2. The Journal of pathology *208*, 507-517.
Zou, F., Chai, H.S., Younkin, C.S., Allen, M., Crook, J., Pankratz, V.S., Carrasquillo, M.M., Rowley, C.N., Nair, A.A., and Middha, S. (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS genetics *8*, e1002707.
Zou, J., Cao, Z., Zhang, J., Chen, T., Yang, S., Huang, Y., Hong, D., Li, Y., Chen, X., and Wang, X. (2016). Variants in human papillomavirus receptor and associated genes are associated with type-specific HPV infection and lesion progression of the cervix. Oncotarget *7*, 40135.